Aus dem Bereich Klinische Bioinformatik
Klinische Medizin
der Medizinischen Fakultät
der Universität des Saarlandes, Homburg/Saar

From tools and databases to clinically relevant applications in miRNA research

Dissertation zur Erlangung des Grades eines Doktors
der Naturwissenschaften der Medizinischen Fakultät

der UNIVERSITÄT DES SAARLANDES
2021

vorgelegt von Tobias Fehlmann
geb. am 07.06.1992 in Emmendingen, Deutschland

Tag der Promotion:	6. Januar 2022
Dekan:	Univ.-Prof. Dr. med. Michael D. Menger
Berichterstatter:	Prof. Andreas Keller
	Prof. Eckart Meese
	Prof. Ulf Leser

"Science is not meant to cure us of mystery, but to reinvent and reinvigorate it." — Robert Sapolsky

Abstract

While especially early research focused on the small portion of the human genome that encodes proteins, it became apparent that molecules responsible for many key functions were also encoded in the remaining regions. Originally, non-coding RNAs, i.e., molecules that are not translated into proteins, were thought to be composed of only two classes (ribosomal RNAs and transfer RNAs). However, starting from the early 1980s many other non-coding RNA classes were discovered. In the past two decades, small non-coding RNAs (sncRNAs) and in particular microRNAs (miRNAs), have become essential molecules in biological and biomedical research.

In this thesis, five aspects of miRNA research have been addressed. Starting from the development of advanced computational software to analyze miRNA data (1), an in-depth understanding of human and non-human miRNAs was generated and databases hosting this knowledge were created (2). In addition, the effects of technological advances were evaluated (3). We also contributed to the understanding on how miRNAs act in an orchestrated manner to target human genes (4). Finally, based on the insights gained from the tools and resources of the mentioned aspects we evaluated the suitability of miRNAs as biomarkers (5).

With the establishment of next-generation sequencing, the primary goal of this thesis was the creation of an advanced bioinformatics analysis pipeline for high-throughput miRNA sequencing data, primarily focused on human. Consequently, miRMaster, a web-based software solution to analyze hundreds sequencing samples within few hours was implemented. The tool was implemented in a way that it could support different sequencing technologies and library preparation techniques. This flexibility allowed miRMaster to build a consequent user-base, resulting in over 120,000 processed samples and 1,5 billion processed reads, as of July 2021, and therefore laid out the basis for the second goal of this thesis. Indeed, the implementation of a feature allowing users to share their uploaded data contributed strongly to the generation of a detailed annotation of the human small non-coding transcriptome. This annotation was integrated into a new miRNA database, miRCarta, modelling thousands of miRNA candidates and corresponding read expression profiles. A subset of these candidates was then evaluated in the context of different diseases and validated. The thereby gained knowledge was subsequently used to validate additional miRNA candidates and to generate an estimate of the number

of miRNAs in human. The large collection of samples, gathered over many years with miRMaster was also integrated into a web server evaluating miRNA arm shifts and switches, miRSwitch. Finally, we published an updated version of miRMaster, expanding its scope to other species and adding additional downstream analysis capabilities.

The second goal of this thesis was further pursued by investigating the distribution of miRNAs across different human tissues and body fluids, as well as the variability of miRNA profiles over the four seasons of the year. Furthermore, small non-coding RNAs in zoo animals were examined and a tissue atlas of small non-coding RNAs for mice was generated.

The third goal, the assessment of technological advances, was addressed by evaluating the new combinatorial probe-anchor synthesis-based sequencing technology published by BGI, analyzing the effect of RNA integrity on sequencing data, analyzing low-input library preparation protocols, and comparing template-switch based library preparation protocols to ligation-based ones. In addition, an antibody-based labeling sequencing chemistry, CoolMPS, was investigated.

Deriving an understanding of the orchestrated regulation by miRNAs, the fourth goal of this thesis, was pursued in a first step by the implementation of a web server visualizing miRNA-gene interaction networks, miRTargetLink. Subsequently, miRPathDB, a database incorporating pathways affected by miRNAs and their targets was implemented, as well as miEAA 2.0, a web server offering quick miRNA set enrichment analyses in over 130,000 categories spanning 10 different species. In addition, miRSNPdb, a database evaluating the effects of single nucleotide polymorphisms and variants in miRNAs or in their target genes was created.

Finally, the fifth goal of the thesis, the evaluation of the suitability of miRNAs as biomarkers for human diseases was tackled by investigating the expression profiles of miRNAs with machine learning. An Alzheimer's disease cohort with over 400 individuals was analyzed, as well as another neurodegenerative disease cohort with multiple time points of Parkinson's disease patients and healthy controls. Furthermore, a lung cancer cohort covering 3,000 individuals was examined to evaluate the suitability of an early detection test. In addition, we evaluated the expression profile changes induced by aging on a cohort of 1,334 healthy individuals and over 3,000 diseased patients.

Altogether, the herein described tools, databases and research papers present valuable advances and insights into the miRNA research field and have been used and cited by the research community over 2,000 times as of July 2021.

Zusammenfassung

Während insbesondere die frühe Genetik-Forschung sich auf den kleinen Teil des menschlichen Genoms konzentrierte, der für Proteine kodiert, wurde deutlich, dass auch in den übrigen Regionen Moleküle kodiert werden, die für viele wichtige Funktionen verantwortlich sind. Ursprünglich ging man davon aus, dass nicht codierende RNAs, d. h. Moleküle, die nicht in Proteine übersetzt werden, nur aus zwei Klassen bestehen (ribosomale RNAs und Transfer-RNAs). Seit den frühen 1980er Jahren wurden jedoch viele andere nicht-kodierende RNA-Klassen entdeckt. In den letzten zwei Jahrzehnten sind kleine nichtcodierende RNAs (sncRNAs) und insbesondere microRNAs (miRNAs) zu wichtigen Molekülen in der biologischen und biomedizinischen Forschung geworden.

In dieser Arbeit werden fünf Aspekte der miRNA-Forschung behandelt. Ausgehend von der Entwicklung fortschrittlicher Computersoftware zur Analyse von miRNA-Daten (1) wurde ein tiefgreifendes Verständnis menschlicher und nicht-menschlicher miRNAs entwickelt und Datenbanken mit diesem Wissen erstellt (2). Darüber hinaus wurden die Auswirkungen des technologischen Fortschritts bewertet (3). Wir haben auch dazu beigetragen, zu verstehen, wie miRNAs koordiniert agieren, um menschliche Gene zu regulieren (4). Schließlich bewerteten wir anhand der Erkenntnisse, die wir mit den Tools und Ressourcen der genannten Aspekte gewonnen hatten, die Eignung von miRNAs als Biomarker (5).

Mit der Etablierung der Sequenzierung der nächsten Generation war das primäre Ziel dieser Arbeit die Schaffung einer fortschrittlichen bioinformatischen Analysepipeline für Hochdurchsatz-MiRNA-Sequenzierungsdaten, die sich in erster Linie auf den Menschen konzentriert. Daher wurde miRMaster, eine webbasierte Softwarelösung zur Analyse von Hunderten von Sequenzierproben innerhalb weniger Stunden, implementiert. Das Tool wurde so implementiert, dass es verschiedene Sequenzierungstechnologien und Bibliotheksvorbereitungstechniken unterstützen kann. Diese Flexibilität ermöglichte es miRMaster, eine konsequente Nutzerbasis aufzubauen, die im Juli 2021 über 120.000 verarbeitete Proben und 1,5 Milliarden verarbeitete Reads umfasste, womit die Grundlage für das zweite Ziel dieser Arbeit geschaffen wurde. Die Implementierung einer Funktion, die es den Nutzern ermöglicht, ihre hochgeladenen Daten mit anderen zu teilen, trug wesentlich zur Erstellung einer detaillierten Annotation des menschlichen kleinen nicht-kodierenden Transkriptoms bei.

Diese Annotation wurde in eine neue miRNA-Datenbank, miRCarta, integriert, die Tausende von miRNA-Kandidaten und entsprechende Expressionsprofile abbildet. Eine Teilmenge dieser Kandidaten wurde dann im Zusammenhang mit verschiedenen Krankheiten bewertet und validiert. Die so gewonnenen Erkenntnisse wurden anschließend genutzt, um weitere miRNA-Kandidaten zu validieren und eine Schätzung der Anzahl der miRNAs im Menschen vorzunehmen. Die große Sammlung von Proben, die über viele Jahre mit miRMaster gesammelt wurde, wurde auch in einen Webserver integriert, der miRNA-Armverschiebungen und -Wechsel auswertet, miRSwitch. Schließlich haben wir eine aktualisierte Version von miRMaster veröffentlicht, die den Anwendungsbereich auf andere Spezies ausweitet und zusätzliche Downstream-Analysefunktionen hinzufügt.

Das zweite Ziel dieser Arbeit wurde weiterverfolgt, indem die Verteilung von miRNAs in verschiedenen menschlichen Geweben und Körperflüssigkeiten sowie die Variabilität der miRNA-Profile über die vier Jahreszeiten hinweg untersucht wurde. Darüber hinaus wurden kleine nichtkodierende RNAs in Zootieren untersucht und ein Gewebeatlas der kleinen nichtkodierenden RNAs für Mäuse erstellt.

Das dritte Ziel, die Einschätzung des technologischen Fortschritts, wurde angegangen, indem die neue kombinatorische Sonden-Anker-Synthese-basierte Sequenzierungstechnologie, die vom BGI veröffentlicht wurde, bewertet wurde, die Auswirkungen der RNA-Integrität auf die Sequenzierungsdaten analysiert wurden, Protokolle für die Bibliotheksvorbereitung mit geringem Input analysiert wurden und Protokolle für die Bibliotheksvorbereitung auf der Basis von Template-Switch mit solchen auf Ligationsbasis verglichen wurden. Darüber hinaus wurde eine auf Antikörpern basierende Labeling-Sequenzierungschemie, CoolMPS, untersucht.

Das vierte Ziel dieser Arbeit, das Verständnis der orchestrierten Regulation durch miRNAs, wurde in einem ersten Schritt durch die Implementierung eines Webservers zur Visualisierung von miRNA-Gen-Interaktionsnetzwerken, miRTargetLink, verfolgt. Anschließend wurde miRPathDB implementiert, eine Datenbank, die von miRNAs und ihren Zielgenen beeinflusste Pfade enthält, sowie miEAA 2.0, ein Webserver, der schnelle miRNA-Anreicherungsanalysen in über 130.000 Kategorien aus 10 verschiedenen Spezies bietet. Darüber hinaus wurde miRSNPdb, eine Datenbank zur Bewertung der Auswirkungen von Einzelnukleotid-Polymorphismen und Varianten in miRNAs oder ihren Zielgenen, erstellt.

Schließlich wurde das fünfte Ziel der Arbeit, die Bewertung der Eignung von miRNAs als Biomarker für menschliche Krankheiten, durch die Untersuchung der Expressionsprofile von miRNAs anhand von maschinellem Lernen angegangen. Eine Alzheimer-Kohorte mit über 400 Personen wurde analysiert, ebenso wie eine weitere neurodegenerative Krankheitskohorte mit Parkinson-Patienten an mehreren Zeitpunkten der Krankheit und gesunden Kontrollen. Außerdem wurde eine Lungenkrebskohorte mit 3.000 Personen untersucht, um die Eignung eines Früherkennungstests zu bewerten. Darüber hinaus haben wir die altersbedingten Veränderungen des Expressionsprofils bei einer Kohorte von 1.334 gesunden Personen und über 3.000 kranken Patienten untersucht.

Insgesamt stellen die hier beschriebenen Tools, Datenbanken und Forschungsarbeiten wertvolle Fortschritte und Erkenntnisse auf dem Gebiet der miRNA-Forschung dar und wurden bis Juli 2021 von der Forschungsgemeinschaft über 2.000 Mal verwendet und zitiert.

Scientific papers

This is a cumulative thesis based on the following published papers. The publications included herein are identical to the published versions. Equally contributing first authors are denoted by a superscript dagger (†) symbol.

[1] **Fehlmann T**, Backes C, Kahraman M, Haas J, Ludwig N, Posch AE, Würstle ML, Hübenthal M, Franke A, Meder B, Meese E, Keller A (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res*, 45:8731–8744

[2] **Fehlmann T**, Kern F, Laham O, Backes C, Solomon J, Hirsch P, Volz C, Müller R, Keller A (2021) miRMaster 2.0: multi-species non-coding RNA sequencing analyses at scale. *Nucleic Acids Res*, 49:W397–W408, W1

[3] Kern F, Amand J, Senatorov I, Isakova A, Backes C, Meese E, Keller A, **Fehlmann T** (2020) miRSwitch: detecting microRNA arm shift and switch events. *Nucleic Acids Res*, 48:W268–W274, W1

[4] **Fehlmann T**†, Backes C†, Alles J, Fischer U, Hart M, Kern F, Langseth H, Rounge T, Umu SU, Kahraman M, Laufer T, Haas J, Staehler C, Ludwig N, Hübenthal M, Meder B, Franke A, Lenhof HP, Meese E, Keller A (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, 34:1621–1628

[5] Backes C, **Fehlmann T**, Kern F, Kehl T, Lenhof HP, Meese E, Keller A (2018) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res*, 46:D160–D167, D1

[6] **Fehlmann T**, Laufer T, Backes C, Kahramann M, Alles J, Fischer U, Minet M, Ludwig N, Kern F, Kehl T, Galata V, Düsterloh A, Schrörs H, Kohlhaas J, Bals R, Huwer H, Geffers L, Krüger R, Balling R, Lenhof HP, Meese E, Keller A (2019) Large-scale validation of miRNAs by disease association, evolutionary conservation and pathway activity. *RNA Biol*, 16:93–103

[7] Alles J†, **Fehlmann T**†, Fischer U, Backes C, Galata V, Minet M, Hart M, Abu-Halima M, Grässer FA, Lenhof HP, Keller A, Meese E (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res*, 47:3353–3364

[8] Ludwig N, Leidinger P, Becker K, Backes C, **Fehlmann T**, Pallasch C, Rheinheimer S, Meder B, Stähler C, Meese E, Keller A (2016)

Distribution of miRNA expression across human tissues. *Nucleic Acids Res*, 44:3865–3877

[9] **Fehlmann T**, Ludwig N, Backes C, Meese E, Keller A (2016) Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol*, 13:1084–1088

[10] Ludwig N, Hecksteden A, Kahraman M, **Fehlmann T**, Laufer T, Kern F, Meyer T, Meese E, Keller A, Backes C (2019) Spring is in the air: seasonal profiles indicate vernal change of miRNA activity. *RNA Biol*, 16:1034–1043

[11] **Fehlmann T**, Backes C, Pirritano M, Laufer T, Galata V, Kern F, Kahraman M, Gasparoni G, Ludwig N, Lenhof HP, Gregersen HA, Francke R, Meese E, Simon M, Keller A (2019) The sncRNA zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Res*, 47:4431–4441

[12] Isakova A, **Fehlmann T**, Keller A, Quake SR (2020) A mouse tissue atlas of small noncoding RNA. *Proc Natl Acad Sci U S A*, 117:25634–25645

[13] **Fehlmann T**, Reinheimer S, Geng C, Su X, Drmanac S, Alexeev A, Zhang C, Backes C, Ludwig N, Hart M, An D, Zhu Z, Xu C, Chen A, Ni M, Liu J, Li Y, Poulter M, Li Y, Stähler C, Drmanac R, Xu X, Meese E, Keller A (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics*, 8:123

[14] Ludwig N†, **Fehlmann T**†, Galata V, Franke A, Backes C, Meese E, Keller A (2018) Small ncRNA-seq results of human tissues: variations depending on sample integrity. *Clin Chem*, 64:1074–1084

[15] Pirritano M†, **Fehlmann T**†, Laufer T, Ludwig N, Gasparoni G, Li Y, Meese E, Keller A, Simon M (2018) Next generation sequencing analysis of total small noncoding RNAs from low input RNA from dried blood sampling. *Anal Chem*, 90:11791–11796

[16] Meistertzheim M†, **Fehlmann T**†, Drews F, Pirritano M, Gasparoni G, Keller A, Simon M (2019) Comparative analysis of biochemical biases by ligation- and template-switch-based small RNA library preparation protocols. *Clin Chem*, 65:1581–1591

[17] Li Y†, **Fehlmann T**†, Borcherding A, Drmanac S, Liu S, Groeger L, Xu C, Callow M, Villarosa C, Jorjorian A, Kern F, Grammes N, Meese E, Jiang H, Drmanac R, Ludwig N, Keller A (2021) CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing. *Nucleic Acids Res*, 49:e10

[18] Hamberg M, Backes C, **Fehlmann T**, Hart M, Meder B, Meese E, Keller A (2016) MiRTargetLink–miRNAs, genes and interaction networks. *Int J Mol Sci*, 17:564

[19] Backes C†, Kehl T†, Stöckel D, **Fehlmann T**, Schneider L, Meese E, Lenhof HP, Keller A (2017) miRPathDB: a new dictionary on microRNAs and target pathways. *Nucleic Acids Res*, 45:D90–D96, D1

[20] Kehl T†, Kern F†, Backes C, **Fehlmann T**, Stöckel D, Meese E, Lenhof HP, Keller A (2020) miRPathDB 2.0: a novel release of the miRNA pathway dictionary database. *Nucleic Acids Res*, 48:D142–D147, D1

[21] **Fehlmann T**†, Sahay S†, Keller A, Backes C (2019) A review of databases predicting the effects of SNPs in miRNA genes or miRNA-binding sites. *Brief Bioinform*, 20:1011–1020

[22] Kern F†, **Fehlmann T**†, Solomon J, Schwed L, Grammes N, Backes C, Van Keuren-Jensen K, Craig DW, Meese E, Keller A (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res*, 48:W521–W528, W1

[23] Ludwig N†, **Fehlmann T**†, Kern F†, Gogol M, Maetzler W, Deutscher S, Gurlit S, Schulte C, von Thaler AK, Deuschle C, Metzger F, Berg D, Suenkel U, Keller V, Backes C, Lenhof HP, Meese E, Keller A (2019) Machine learning to detect alzheimer's disease from circulating non-coding RNAs. *Genomics Proteomics Bioinformatics*, 17:430–440

[24] Kern F, **Fehlmann T**, Violich I, Alsop E, Hutchins E, Kahraman M, Grammes NL, Guimarães P, Backes C, Poston KL, Casey B, Balling R, Geffers L, Krüger R, Galasko D, Mollenhauer B, Meese E, Wyss-Coray T, Craig DW, Van Keuren-Jensen K, Keller A (2021) Deep sequencing of sncRNAs reveals hallmarks and regulatory modules of the transcriptome during parkinson's disease progression. *Nat Aging*, 1:309–322

[25] **Fehlmann T**†, Kahraman M†, Ludwig N, Backes C, Galata V, Keller V, Geffers L, Mercaldo N, Hornung D, Weis T, Kayvanpour E, Abu-Halima M, Deuschle C, Schulte C, Suenkel U, von Thaler AK, Maetzler W, Herr C, Fähndrich S, Vogelmeier C, Guimaraes P, Hecksteden A, Meyer T, Metzger F, Diener C, Deutscher S, Abdul-Khaliq H, Stehle I, Haeusler S, Meiser A, Groesdonk HV, Volk T, Lenhof HP, Katus H, Balling R, Meder B, Kruger R, Huwer H, Bals R, Meese E, Keller A (2020) Evaluating the use of circulating MicroRNA profiles for lung cancer detection in symptomatic patients. *JAMA Oncol*, 6:714–723

[26] **Fehlmann T**, Lehallier B, Schaum N, Hahn O, Kahraman M, Li Y, Grammes N, Geffers L, Backes C, Balling R, Kern F, Krüger R, Lammert F, Ludwig N, Meder B, Fromm B, Maetzler W, Berg D, Brockmann K, Deuschle C, von Thaler AK, Eschweiler GW, Milman S, Barziliai N, Reichert M, Wyss-Coray T, Meese E, Keller A (2020) Common diseases alter the physiological age-related blood microRNA profile. *Nat Commun*, 11:5958

Contents

List of Figures

List of Tables

Abbreviations

A

AD Alzheimer's disease 29–31

AGO Argonaute 36

ANOVA analysis of variance 48

AUC area under the curve 49

AUC-ROC area under the curve of the receiver operating characteristic curve 50

C

C. elegans *Caenorhabditis elegans* 33, 34, 51

cDNA complementary DNA 41–44

circRNA circular RNA 37

COVID-19 coronavirus disease 2019 59

CSS Cascading Style Sheets 55

CT computed tomography 28

D

DGCR8 DiGeorge Syndrome Critical Region 8 36

DNA deoxyribonucleic acid 29, 32, 41–44

dNTP deoxynucleotide triphosphates 41, 44

DOM Document Object Model 55

E

ENCODE Encyclopedia of DNA Elements 55

G

GEO Gene Expression Omnibus 55

GO Gene Ontology 54

GSEA Gene Set Enrichment Analysis 54

GTEx Genotype-Tissue Expression 55

H

H. sapiens *Homo sapiens* 34
HMDD Human MicroRNA Disease Database 54
HTML HyperText Markup Language 55

K

KEGG Kyoto Encyclopedia of Genes and Genomes 54

L

lncRNA long non-coding RNA 33

M

M. musculus *Mus musculus* 34
MDS International Parkinson and Movement Disorder Society 31
miRNA microRNA
MRE miRNA response element 36–38
mRNA messenger RNA 33, 36–39, 42, 53, 333, 334

N

ncRNA non-coding RNA 32
NGS next-generation sequencing 34, 35, 38, 39, 43, 45–47, 49, 51, 52, 57, 332
NSCLC non-small cell lung cancer 28, 29
nt nucleotide 33–36

P

PABPC cytosolic poly(A) binding protein 36
PCA principal component analysis 47, 48
PCR polymerase chain reaction 41–43, 45
PD Parkinson's disease 30, 31
PDB Protein Data Bank 55
PET positron emission tomography 29–31
piRNA Piwi-interacting RNA 33, 44
pre-miRNA precursor miRNA 36, 51, 52

R

RFE recursive feature elimination 50
RIN RNA integrity number 41
RISC RNA-induced silencing complex 33, 36, 39
RNA ribonucleic acid
RNAi RNA interference 33
RPM reads per million 45
RPMM reads per million mapped 45
RPMMM reads per million mapped to miRNAs 45
rRNA ribosomal RNA 32
RT-qPCR quantitative reverse transcriptase polymerase chain reaction 38–43, 45, 49

S

SCLC small cell lung cancer 28
SEER Surveillance, Epidemiology, and End Results 28
siRNA small interfering RNA 33
SLAM-seq thiol(SH)-linked alkylation for the metabolic sequencing 334
sncRNA small non-coding RNA 32, 41, 43–49, 57
snoRNA small nucleolar RNA 36, 42
snRNA small nuclear RNA 42
SNV single nucleotide variant 53, 58, 334
SRA Sequence Read Archive 55

T

t-SNE t-distributed stochastic neighbor embedding 47, 48
TCGA The Cancer Genome Atlas 55
tRNA transfer RNA 32, 33, 36
tsRNA tRNA-derived small RNA 32, 33

U

UICC Union for International Cancer Control 28
UMAP uniform manifold approximation and projection 47, 48
UMI unique molecular identifier 45
UTR untranslated region 33, 37

W

WHO World Health Organization 27

1 microRNAs: tiny molecules with large impact

1.1 Motivation

Many widespread diseases could benefit from earlier detection. Small non-coding ribonucleic acids (RNAs), especially microRNAs, are molecules that have the potential to contribute to the early detection of those diseases.

1.1.1 Current disease burden

Since the start of the last century the worldwide average life expectancy has more than doubled, reaching around 80 years in high-income countries [27]. At the same time medical treatment and prevention of previously major death factors, such as influenza, tuberculosis or gastrointestinal infections have become possible [28], thereby increasing the importance of healthy aging. These changes are reflected in the current mortality statistics and especially in high-income countries. According to the World Health Organization (WHO), the main causes of death in high-income countries are ischaemic heart disease, Alzheimer's disease and other dementias, stroke, and trachea, bronchus and lung cancers (see Table 1.1, [29]). In particular, aging related neurodegenerative diseases are on the rise. With an increase from 2.4% of all deaths in 2000 to 7.6% in 2019, Alzheimer's disease and other dementias rapidly affect an increasingly large population. While currently around 50 million people have dementia, it is expected that the number of cases will continue to grow, with projected 82 million in 2030 and 152 million in 2050 [30]. A similar trend can be observed for Parkinson's disease, the second most frequent neurodegenerative disorder, for which the death rates increased from 0.5% in 2000 to 1.1% in 2019, thus ranking as the 20th most frequent cause of death in high-income countries. In 2016, about 6.1 million individuals were affected by Parkinson's disease worldwide and data from the USA suggest a 50% increase by 2032 [31]. While lung cancer is still the deadliest cancer in high-income countries, its death rates only changed marginally over the last 20 years, from 5.5% of all deaths in 2000 to slightly lower 5.4% in 2019. Global statistics however show that lung

Table 1.1: Top 10 causes of death in 2000 and 2019 in high-income countries, according to the World Health Organisation.

Cause	**Deaths in 2019**	**% of total deaths in 2019**	**% of total deaths in 2000**
Ischaemic heart disease	1,729K	16.1	22.5
Alzheimer disease and other dementias	814K	7.6	2.4
Stroke	790K	7.4	10.9
Trachea, bronchus, lung cancers	584K	5.4	5.5
Chronic obstructive pulmonary disease	551K	5.1	4.6
Lower respiratory infections	419K	3.9	4.1
Colon and rectum cancers	331K	3.1	3.2
Kidney diseases	270K	2.5	1.8
Hypertensive heart disease	220K	2.1	1.3
Diabetes mellitus	202K	1.9	2.0

cancer starts to affect lower-income countries more severely, with an overall change of 2.4% of all deaths in 2000 to 3.2% in 2019 [29]. For many of the mentioned diseases an early diagnosis is key to improve the outcome and increase the quality of life.

1.1.2 *Lung cancer*

Lung cancer is defined by a malignant tumor located in the lung [32]. A tumor is characterized by the abnormal growth of cells. The progress of the disease is grouped into stages defined by the Union for International Cancer Control (UICC), relative to the spread of the tumor [33]. Two major disease types are distinguished: non-small cell lung cancer (NSCLC), accounting for about 85% of all cases, and small cell lung cancer (SCLC) [34]. At the molecular level, lung cancer is involved in a multitude of hallmarks of cancer. It is characterized by the expansion into neighboring tissues as well as more distant locations, the evasion of cell death, self-sufficient proliferation, escaping of growth suppressors, induction of blood vessel growth, and unlimited replication potential, as well as the escaping of immune destruction and deregulation of the metabolism [35, 36]. Patients affected by lung cancer show nearly no particular signs in early stages, while in more advanced stages they present respiratory symptoms like coughing, systemic symptoms like weight loss, and symptoms due to the expansion of the cancer like chest and bone pain [32]. The largest risk factor for lung cancer is by far smoking, which is estimated to account of 90% of all cases [37, ch. 2]. The early diagnosis of lung cancer is particularly important since the National Cancer Institute in the USA reports in its Surveillance, Epidemiology, and End Results (SEER) program a 5-year survival rate of 60% for localized tumors (UICC stage 1), 33% for regional tumors (stage 2) and 6% for distant tumors (stage 3 and 4) [38]. However, only 18% of the patients were diagnosed with localized tumors, while 56% of the patients already had distant tumors, thus greatly reducing the overall survival rate. Although it resulted in a high false-positive rate, low-dose computed tomography (CT) screening was identified by the National Lung Cancer Screening Trial in the USA as potential early detection method that could reduce overall lung cancer mortality by 20% [39]. Other early detection strategies that are currently explored include, among others, molecular measurements in the breath condensate [40, 41], and blood-borne biomarkers measuring protein or microRNA levels in whole-blood, serum, or plasma [25, 42–44], but also measurements of circulating tumor deoxyribonucleic acid (DNA) [45, 46].

While the early detection is a key factor for an efficient and effective treatment, the discovery of new therapeutic targets to treat diseases is central to benefit from said detection. With the rise of precision medicine, many medical advances found their way into the clinic in the last 10 years, especially for NSCLC patients. The use of video-assisted thoracic surgery resulted in lower perioperative morbidity [47, 48] and the application of stereotactic ablative radiotherapy offered a promising alternative to early-stage surgery [49, 50]. As alternative, or complementary to chemotherapy, the identification of a wide set of molecular oncogenic drivers led to an increase in targeted therapies [51–53]; and the development of immunotherapy by identification of immune checkpoint inhibitors presented complementary or alternative treatments [54–56].

1.1.3 Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disease for which age represents the major risk factor, mostly affecting individuals over the age of 65. The accumulation of extracellular β-amyloid (Aβ)-containing plaques in the brain is a characteristic of the neuropathology of the disease, as well as deposits of intracellular neurofibrillary tangles and brain atrophy [57]. At the molecular level, AD is characterized by various hallmarks. One major hallmark is the misfolding and aggregation of the Aβ peptides, produced from the enzymatic cleavage of the amyloid precursor protein [58]. The next is the aggregation of hyperphosphorylated microtubule associated protein, known as tau, resulting from different post-translational modifications [59]. Additional cofactor hallmarks are oxidative stress [60], mitochondrial dysfunction [60] as well as autophagy dysfunction [61], endoplasmic reticulum stress [62] and inflammation [63]. The diagnosis of the disease itself is still evolving [64]. Particularly, before the last decade AD was diagnosed mostly through symptomatic signs, after the identification of a memory related cognitive disorder that would hamper the daily routine of the patient and exclusion of other possibilities, resulting in the labeling as "probable AD". Definite AD was only diagnosed postmortem, after the identification of Aβ plaques and neurofibrillary tangles in the brain. Nowadays, AD is diagnosed through a combination of clinical criteria and biomarkers for Aβ and tau levels in cerebrospinal fluid and the brain through positron emission tomography (PET) [65]. Overall, it is accepted that the disease progresses through three major stages (sometimes also split into five or seven), starting with the preclinical stage with small brain changes, followed by mild cognitive impairment and ending with dementia [66–68]. Early diagnosis for AD is especially difficult because the signs at the preclinical stage and early mild cognitive impairment are often dismissed as being related to the age of the individual. Although no cure for AD is available yet, an early diagnosis provides the possibility to lessen the symptoms and potentially delay the progression of the disease, thereby leading to an increased quality of life. However, no established early diagnosis method is available yet. Currently evaluated approaches include measuring tau, Aβ, as well as neurofilament light chain, and neurogranin levels in cerebrospinal fluid or plasma [69–76], as well as measuring microRNA levels in whole-blood, serum, or plasma [77–79]. Additionally, brain imaging techniques are considered to measure tau and Aβ levels, as well as general markers of neurodegeneration. Among these techniques, one is structural imaging that allows to characterize the shrinkage of the brain, via e.g., magnetic resonance imaging [65]. Another is functional imaging, which allows to indirectly investigate the neuronal activity via PET [65]. Finally, molecular imaging allowing to detect Aβ [80, 81] and tau [82] in the brain is evaluated. In contrast to the breakthroughs made in lung cancer treatment, new treatments for Alzheimer's disease are progressing more slowly. Current developments try to reduce or remove the Aβ plaques [83] or prevent synapse destruction [84]. Other approaches involve the prevention of tau tangling [85], the reduction of neuroinflammation [86] and the enhancement of cognition via plasma fractions [87].

1.1.4 Parkinson's disease

Parkinson's disease (PD) is a neurodegenerative disease mostly affecting individuals over the age of 60 [88]. Although PD can be inherited, it only represents 5-10% of all cases [89]. Its neuropathology is described by the loss of dopaminergic neurons in certain areas of the substantia nigra and intracellular α-synuclein protein accumulation. Its molecular hallmarks include the misfolding and aggregation of α-synuclein proteins, lysosomal and proteasomal dysfunction, mitochondrial dysfunction, oxidative stress, disruption of calcium homeostasis, neuroinflammation, neuron death as well as synaptic dysfunction [89, 90]. As for AD, the diagnosis of PD is continuously evolving. It is currently diagnosed first by the diagnosis of parkinsonism, characterized among other by a movement disorder, and second by at least two supportive criteria like, positive response to dopaminergic therapy, rest tremor, Levodopa-induced dyskinesia, olfactory loss, or cardiac sympathetic denervation, and by the absence of absolute and unusual exclusion criteria [91]. The disease is labeled as "probable PD" if unusual exclusion criteria that are counterbalanced by supportive criteria are present. Like AD, PD is characterized by multiple stages, starting with prodromal PD with non-motor features, followed by early-stage PD including motor features and often mild cognitive impairment, progressing through mid-stage PD with stronger motor features and additional non-motor symptoms like orthostatic hypotension, and ending late-stage PD with even stronger motor features and dementia [89]. Although no cure for PD yet exists, early diagnosis of PD is important since available treatments allow to greatly improve the quality of life. Currently evaluated diagnostic biomarkers include diverse brain imaging techniques monitoring e.g., cardiac sympathetic denervation [92], integrity of pigmented neurons of specific brain regions [93] or brain function [94]. In addition, various tissue biopsies of e.g., the skin and submandibular gland show promising results [95, 96]. Furthermore, continuous monitoring through smart wearables is considered [97]. Similarly to AD, measurements of protein levels from cerebrospinal fluid and blood, plasma, or serum [98, 99] as well as other molecular markers such as microRNAs (miRNAs) are evaluated for early detection, as well as progression [24, 100–102]. Official markers selected by the International Parkinson and Movement Disorder Society (MDS) to be included into a Bayesian classifier identifying prodromal cases include abnormal Dopaminergic PET/SPECT, polysomnography-proven rapid eye movement sleep behavior disorder, olfactory dysfunction and orthostatic hypotension, as well as subthreshold parkinsonism, among others [103]. Treatments for PD currently use drugs to restore the dopamine levels via Levodopa-based drugs (converted into dopamine in the brain), dopamine agonists that mimic dopamine effects in the brain and inhibitors of enzymes that metabolize dopamine [89]. Other drugs that don't target the dopamine loss are also employed and developed focused on treating parkinsonism, tremor, and gait disorders, but also to counter Levodopa-induced dyskinesia [89]. In addition, deep brain stimulation, i.e., electric stimulation of specific brain region through implanted electrodes, is used to reduce PD symptoms [104]. Furthermore, gene therapy approaches [105], cell transplantations [106], as well as immunotherapeutics [107] are investigated.

1.1.5 *Aging*

With aging being the highest risk factors for many diseases, healthy aging is of utmost importance. Aging affects our whole body and is manifested through various changes at physiological and molecular level. In 2013, López-Otín *et al.* proposed nine hallmarks central to mammalian aging and grouped them into three categories. The first category is composed of primary hallmarks, which are determined as the causes of damage, "genomic instability, telomere attrition, epigenetic alterations and loss of proteostasis". The second is composed of antagonistic hallmarks, which are responses to damage: "deregulated nutrient sensing, mitochondrial dysfunction and cellular senescence". The last category consists of integrative hallmarks, which affect the phenotype through tissue and function: "stem cell exhaustion and altered intercellular communication" [108]. Still, the mechanisms of these processes, their interactions and occurrences need to be further researched. To this end various molecular studies have been performed, focusing on multiple aspects of aging with different omics technologies. Longevity, healthspan and lifespan related genes were investigated by genome-wide association studies [109], plasma protein signatures of aging were examined [110], different aging patterns of organs were derived from single-cell and bulk RNA sequencing [111, 112], the importance of DNA methylation to derive an epigenetic aging clock was evaluated [113], and microRNA blood signatures were tested as biomarkers for accelerated aging [114].

Different strategies to reduce, counter, or even inverse the effects of aging have been pursued in mammals and include dietary restriction [115], telomerase manipulations [116], nuclear reprogramming [117], as well as young blood injections [118].

1.2 Biological background

One of the key aspects of this thesis is the use of circulating small non-coding RNAs, in particular miRNAs, to discover biomarkers for lung cancer, Alzheimer's disease, and Parkinson's disease patients, as well as generating an understanding of molecular healthy and pathological aging. To this end it is essential to understand what characteristics define a miRNA, how it functions and what technical solutions and limitations exist to measure it.

1.2.1 Small non-coding RNAs

Non-coding RNAs (ncRNAs) are RNA molecules that are not translated into proteins. They can be grouped by length into either small non-coding RNAs (sncRNAs), which are typically between 18 and 200 nucleotides (nt) long, or longer RNAs. Non-coding RNAs are composed of functionally different RNA classes, such as ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) that are essential for protein synthesis, amongst others, as represented in Table 1.2. One of the most recently discovered sncRNA classes are tRNA-derived small RNAs (tsRNAs), which were identified in 2008 [119] and since have been shown to be involved in many biological processes [120–122].

Table 1.2: Overview of the most important classes of eukaryotic non-coding RNAs (ncRNAs). Year of discovery indicates the year a class was first described in the literature.
* PTR: post-transcriptional regulation.

RNA	Full name	Length (nt)	Function	Year of discovery	References
rRNA	ribosomal RNA	1800-5000	Protein synthesis	1955	[123]
tRNA	transfer RNA	76-90	Protein synthesis	1957	[124]
snRNA	small nuclear RNA	~150	Processing of nuclear pre-mRNAs	1968	[125]
snoRNA	small nucleolar RNA	60-300	Guides for chemical modifications of other RNAs	1976	[126]
Y RNA	Y RNA	80-120	DNA replication, Ro60 inhibition	1981	[127, 128]
lncRNA	long non-coding RNA	>200	Diverse (regulation of transcription, PTR*, genome integrity, structural functions, etc.)	1984	[129]
miRNA	microRNA	21-25	RNA interference	1993	[130]
siRNA	small interfering RNA	20-27	RNA interference	1998	[131]
scaRNA	small Cajal body-specific RNA	60-300	Guides for chemical modifications of other RNAs	2001	[132]
piRNA	Piwi-interacting RNA	26-31	Gene expression regulation, transposon defense	2004	[133]
tsRNA	tRNA derived small RNA	16-40	Diverse (regulation of transcription, PTR*, cell proliferation, tumor genesis, stress response, etc.)	2008	[119]

Most common RNA classes can be found in Rfam, a database collecting RNA families created through the integration of sequences of many species [134]. Class-specific non-coding RNAs are often stored in specialized databases, such as GtRNADb [135] (for tRNAs), piRBase [136] (for Piwi-interacting RNAs (piRNAs)), miRBase [137] (for miRNAs), but also NONCODE [138] (for long non-coding RNAs (lncRNAs)), and tsRBase [139] (for tsRNAs), which are largely integrated into RNAcentral [140].

1.2.2 *RNA interference*

RNA interference (RNAi) is a process involved in the regulation of gene expression. It is characterized by the RNA-induced silencing complex (RISC) and the binding of complementary RNA molecules to target messenger RNA (mRNA) transcripts. The process of transcriptional repression by antisense RNA was originally observed in transgenic plants [141] and further investigated in other organisms [142, 143]. In 1998, Fire and Mellow described the RNAi mechanism in *Caenorhabditis elegans* (*C. elegans*) [144], for which they won the Nobel Prize in Physiology or Medicine in 2006. Since then, it was found that RNAi is well conserved in many eukaryotes [145–149]. The RNAi mechanism is based on the processing of double-stranded precursor RNA molecules, which can be grouped into two different small non-coding RNA classes, small interfering RNAs (siRNAs) and microRNAs (miRNAs). siRNAs are short endogenous or exogenous double-stranded molecules of about 20-27 base pairs, while miRNAs are endogenous short single-stranded molecules of usually 22 nucleotides. Those molecules are further processed and guide the RISC to the target sequence. It was shown that RNAi happens at the post-transcriptional level and leads to mRNA cleavage or translational repression. Since siRNAs act by perfect complementary binding to their target mRNA, synthetic siRNAs can be used to knock down any protein. This was used to study various biological questions, mainly focused on the functional analysis of genes [150–154], but also evaluated for biomedical applications [155–159]. However, the intracellular delivery of siRNAs is challenging, in particular because of off-targeting induced side-effects [160] and immune responses [161]. In this thesis the focus will be placed on miRNAs, the other class of small non-coding RNAs that participate in RNAi, which will be detailed in the following sub-section.

1.2.3 *microRNAs*

The establishment of microRNAs The first miRNA was discovered in 1993 by Lee, Feinbaum and Ambros [162], while studying the *lin-4* gene in *C. elegans* and its regulatory relationship with *lin-14*. The *lin-4* gene is involved in the regulation of the developmental timing of *C. elegans*. They identified the *lin-4* miRNA with 22 nucleotides (nts) and another transcript of 61 nts, which defined the precursor miRNA. In addition, they found that short regions of the *lin-4* miRNA were complementary to sequences in the 3′ untranslated region (UTR) of the *lin-14* mRNA. At this time, it was assumed that the regulation happened via antisense RNA-RNA interaction, since RNAi was not known then. Seven years later, in 2000, the second miRNA was discovered by Reinhart *et al.* [163], while studying another gene involved in *C. elegans* development, the *let-7* gene. The same year, the 21 nts long *let-7* miRNA was identified in 14 other bilaterian animals [164], among which were *Homo sapiens* (*H. sapiens*) and *Mus musculus* (*M. musculus*), confirming that these small RNAs were not specific to worms but likely part of a class of small RNAs. In the following year, 2001, several RNAs similar to the *lin-4* and *lin-7* miRNAs were discovered [165–167], and the term microRNAs (miRNAs) was established. Since then, the interest in miRNAs has grown exponentially (with nearly 140,000 publications from 2001 until 2020, see Figure 1.1). One causal factor is that miRNAs are expected to regulate a large majority of biological and cellular processes [168], given their ability to regulate a broad set of target genes. In particular, miRNAs have been shown to play a central role in the orchestrated regulation of whole pathways [169]. Unsurprisingly, abnormal miRNAs levels have been linked with a large panel of disease conditions [170–173].

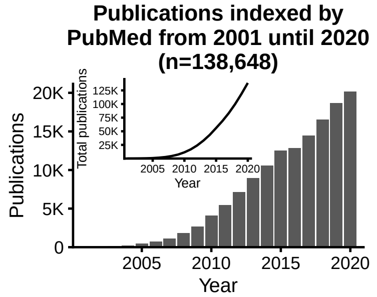


Figure 1.1: PubMed indexed publications containing the search term "miRNA" or "microRNA" shown per year as bar plot and total number of articles shown as line plot.

Reference database In 2004, the first publication of miRBase (then known as "the miRNA Registry"), the current reference database, was released with 506 miRNAs in 6 species [174]. With the establishment of next-generation sequencing (NGS), the number of identified miRNAs grew quickly, covering as of today 48,860 miRNAs in 271 species in the latest release of miRBase v22 [137]. Its annotation system was described in 2003 by Ambros *et al.* [175]. Briefly, miRNAs are numbered consecutively and start with a three-letter code to describe the organism, followed by "mir" to denote the miRNA gene or stem-loop and end with a numeric identifier (e.g. hsa-mir-17 for the miRNA gene). The mature miRNAs derived from a miRNA gene follow a similar convention. Instead of the lower case "mir", they are denoted by "miR" and are followed by 5p or 3p, depending on their proximity to the 5′ or 3′ end of the stem-loop, when the origin of the mature sequence is known. In case a precursor sequence generates the same mature sequences at different genomic loci, those precursors are suffixed by an additional numeric identifier (e.g., hsa-mir-24-1 and hsa-mir-24-2). Closely related mature sequences are additionally distinguished by sharing the same numerical identifier suffixed by a letter (e.g., hsa-miR-19a-3p and hsa-miR-19b-3p). In addition, identical miRNAs in different species usually share the same numeric identifier. Exceptions to these rules exist, such as the first discovered miRNA genes let-7 and lin-4, but also plant and viral miRNAs. Because miRNAs and their precursors are evolutionary conserved, and their naming not always allows to infer their relationship, precursor families were introduced [176]. This was particularly important after the naming conventions since the first release of miRBase changed. One of strongest changes affected the original distinction between the predominant miRNA sequence (also known as "major" miRNA, e.g., hsa-miR-19) and less strongly expressed sequence (also known as "minor" miRNA or "star" sequence, e.g., hsa-miR-19a*), which was replaced by the arm suffix notation (e.g., hsa-miR-19a → hsa-miR-19a-3p and hsa-miR-19a* → hsa-miR-19a-5p). This change was introduced with the release of miRBase v20 [177] and was in part a consequence of many studies recognizing that the observed dominant form was dependent on different factors such as the originating tissue, cell type, or disease state [178–181]. Changes from one dominant arm to another were termed "arm switch". As a result, miRBase name trackers and converters were established, to enable joint analysis of data stemming from different versions [182, 183]. After numerous publications criticizing the number of false positive entries [184–188], miRBase introduced an additional annotation for its miRNAs and stem-loops, allowing them to distinguish between "high-confidence" and "low-confidence" miRNAs [137, 177]. To this end they required (1) the stem-loops to have two annotated miRNA sequences, (2) the miRNA duplex to have a 3′ overhang between 0 and 4 nts, (3) evidence of at least 20 sequencing reads for each miRNA and (4) at least 50% of the reads of one miRNA to start at the same 5′ position.

Alternative databases Although miRBase is the *de facto* standard miRNA reference database, criticism regarding false positive entries and inconsistent naming schemes across organisms gave rise to another miRNA database called miRGeneDB, which started as a curated subset of miRBase in 2015 [189]. This database was updated in 2019, and now contains 10,899 miRNAs from 45 species, with 7 species not represented in miRBase [190]. Its goal is to present a curated established set of miRNAs, with more accurate canonical miRNA sequences and distinguished mature and star sequences, as well as a naming scheme derived from an evolutionary point of view.

Because of an increasing number of studies reporting up to thousands of miRNA candidates in human [184, 191–193], and the exponential grow of available NGS datasets [194, 195], a new database collecting these candidates was established and named miRCarta [5]. The goal of this database is to address the issue of repeated discoveries of miRNA candidates by presenting a sensitive reference resource and to provide expression pattern information for these molecules.

In addition, two other specialized databases were published recently. PmiREN focuses on plant miRNAs that were manually curated after application of an NGS miRNA prediction pipeline using miRDeep-P2 [196]. In addition, a database focusing exclusively on mirtrons (miRNAs derived from introns) collected from the literature, named mirtronDB was published [197].

Biogenesis A multitude of pathways lead to the biogenesis of miRNAs, typically grouped into canonical and noncanonical pathways. While these pathways are very similar among metazoans, the biogenesis of miRNAs is triggered by different pathways in other organisms such as plants and fungi [198–200]. The following sections will cover the mechanisms known in metazoans.

The canonical miRNA biogenesis starts with the transcription of a long RNA, termed pri-miRNA [201, 202]. From this RNA, one miRNA or sometimes multiple miRNAs (such as the miR-17/92 cluster) will be produced [203]. At least one local stem-loop structure is formed and then cleaved by the microprocessor complex [204, 205], composed of an endonuclease with two ribonuclease III domains, Drosha, and the cofactor responsible for its activation, the protein DiGeorge Syndrome Critical Region 8 (DGCR8). This results in the precursor miRNA (pre-miRNA), a shorter hairpin of approximately 60-70 nucleotides with a characteristic 2-nt overhang [206]. The pre-miRNA is then exported to the cytoplasm via Exportin 5 and RanGTP [207, 208], where it is further processed by Dicer, another endonuclease with two ribonuclease III domains [209–211]. The hairpin is cleaved near the loop region, again with a characteristic 2-nt overhang [212], resulting in a miRNA duplex, composed of two mature miRNAs. The duplex is loaded into a protein of the Argonaute (AGO) family via an ATP involved process assisted by the Hsc70/Hsp90 chaperone machinery [213], after which only the guide strand remains loaded, while the passenger strand is separated [214]. Which miRNA will be eventually loaded depends on a multitude of factors such as the cell type, developmental stage or disease [215], but also on 5′-end nucleotides, as well as the thermodynamic stability. Strands with a 5′-uridine or 5′-adenosine are preferentially selected [216, 217], as well as strands with the thermodynamically least stable 5′ end [218, 219]. The complex formed by the AGO protein and the miRNA, known as RISC, is then guided by the miRNA to target transcripts. If sufficiently strong pairings occur, the target is degraded or its translation is repressed [220], with degradation being the most common case in metazoa [221, 222]. The mRNA degradation mechanism starts with the recruitment of TNRC6 by AGO [223, 224], which acts together with the cytosolic poly(A) binding protein (PABPC), leading to the recruitment of PAN2-PAN3 and CCR4-NOT complexes [225–227]. These complexes result in the deadenylation of the poly(A)-tail of the mRNA, which in turn promotes the decapping of the mRNA [226] leading to its degradation by XRN1 [228].

In addition to the afore-described biogenesis pathway, several other processes were shown to produce miRNAs. A Drosha independent biogenesis pathway was found for miRNAs resulting from pre-miRNAs that derived from spliced and debranched introns known as mirtrons [229]. In addition, miRNAs bypassing Drosha can also be generated from endogenous short-hairpin RNAs [230]. Another pathway is deriving miRNAs from other non-coding RNAs like small nucleolar RNAs (snoRNAs) [231] or tRNAs [230], thereby also bypassing Drosha. A Dicer independent pathway was found as well, which starts with the canonical Drosha cleavage, but then loads the precursor miRNA directly into the AGO protein, where it is cleaved and further trimmed at the 3′ end by the poly(A)-specific ribonuclease [232, 233].

Target site characteristics Translational repression or degradation of a target mRNA can only happen if the miRNA can bind sufficiently well to the target mRNA. Target regions, also known as miRNA response elements (MREs), are usually located in the 3′ UTR of the mRNAs, although they have also been reported to occur in 5′ UTRs and open reading frames [234–236]. MREs can also be located in a class of RNAs called circular RNAs (circRNAs), that can act as so-called miRNA sponges which inhibit the binding to other targets [237]. The nucleotides 2 to 7 of the miRNA, which are known as the "seed", play a central role in the recognition and complementary binding to MREs [238]. This region is highly conserved in mammals [239], as are also the MREs [234]. Additional variations known to influence the binding are pairings from position 3 to 8, 2 to 7 with an adenosine at the first position in the MRE, 2 to 8 and 2 to 8 with an adenosine at the first position in the MRE [168, 234, 240], as depicted in Figure 1.2 . MREs where pairings occur only through the seed region are known as canonical sites. Moreover, some sites benefit from additional pairing of nucleotides 13 to 16 [240–242]. Noncanonical sites are characterized by mismatches in the seed region, or shorter seeds of only 5 nucleotides, which can be compensated by more extensive pairing in the 3′ region of the miRNA [238, 243–246]. Overall, these binding patterns result in the possibility of one miRNA regulating multiple genes, as well as many miRNAs possibly regulating the same gene.

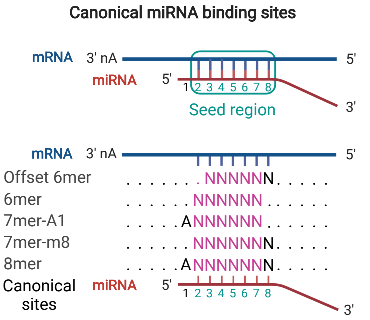

Canonical miRNA binding sites
mRNA
3' nA
5'
miRNA
5'
1 2 3 4 5 6 7 8
Seed region
3'
mRNA
3' nA
5'
Offset 6mer
. . . . NNNNNN. . . . .
6mer
. . . . . . NNNNNN . . . . . .
7mer-A1
. . . . . ANNNNNN . . . . . .
7mer-m8
. . . . . . NNNNNNN. . . . .
8mer
. . . . . ANNNNNNN. . . . .
Canonical sites
miRNA
5'
1 2 3 4 5 6 7 8
3'

Figure 1.2: Canonical miRNA binding sites. The seed region is colored in teal and the position of the canonical seed is colored in purple. The terminology used for the binding site types is identical to the one in [247]. Created with BioRender.com.

Sequence modifications Modifications affecting the sequence of a miRNA, especially if they happen in the seed region, are expected to affect the target transcriptome. In addition, such changes can affect miRNA biogenesis by altering the RNA secondary structure. Modified miRNA sequences that differ in length from the reference sequence or that are affected by nucleotide changes are known as isomiRs [248]. They can exhibit longer or shorter 5′ ends, longer or shorter 3′ ends, or both, as well as edited nucleotides [4, 249–253]. Their prevalence has been found to partially depend on tissue and cell type [250, 254, 255], developmental stage [256], chromosomal sex [257], as well as disease state [258, 259]. Moreover, some miRNAs seem to be more likely to produce isomiRs than others [260]. Changes to the 5′ end and 3′ end are triggered by alternative cleavage by Drosha or Dicer, depending on the location of the miRNA in its precursor [251, 261, 262]. Variations at the 5′ end result in a different seed sequence, thereby greatly affecting the potential target sites [263–265]. Modifications to the 3′ end can induce changes to the miRNA stability and turnover [266, 267], and be triggered by non-templated nucleotide additions. Non-templated nucleotide additions are mostly accomplished through adenylation or uridylation [268, 269]. The miRNA sequence itself can change through RNA editing, with the adenosine to inosine conversion catalyzed by the deaminase ADAR being the most commonly observed [252, 270]. RNA editing usually happens at the miRNA precursor level and can result in reduced miRNA expression by interfering with Drosha [270, 271] or Dicer [272], but also, if affecting the seed region, result in different target sites [273–275]. IsomiRs are defined relative to a reference sequence, known as the canonical miRNA sequence. However, since miRNAs in miRBase are derived from publications where they were originally discovered and because their prevalence depends on multiple factors, the reference sequence reported by miRBase is not necessarily the most frequently encountered form and can actually represent a rarely occurring isoform.

Single nucleotide variants Single nucleotide variants and single nucleotide polymorphisms present in the genome are known to play a role in many diseases, such as Alzheimer's disease [276] and multiple cancers [277]. In the last two decades the number of known variants has grown tremendously, encompassing over 500 million variants currently stored in dbSNP [278]. Such variants, in particular if located in coding regions, might affect protein folding, binding, and expression. However, variants can also occur in miRNAs, as well as in their target genes. These can have as consequence the gain or loss of new MREs and thus affect the regulatory landscape of miRNAs and mRNAs. This is reflected in studies finding effects in schizophrenia [279], progressive hearing loss [280], as well as in many cancers [281–283].

miRNA candidate validation With the ever-rising numbers of discovered potential miRNAs, mainly through *in silico* methods, experimental validation of these sequences is key before they can be considered true miRNAs. Many different validation methods exist, focusing on different aspects of miRNA biogenesis or expression. Some methods use quantitative reverse transcriptase polymerase chain reaction (RT-qPCR) or microarrays to confirm that putative miRNAs are expressed [284]. This has the disadvantage that the possibility of those sequences being other RNAs still cannot be excluded. Others use knockout systems for Drosha and/or Dicer followed by NGS or RT-qPCR quantification to show the presence or absence of the processed sequences [285, 286]. This method has the intrinsic drawback that the modifications of Drosha or Dicer are expected to be the cause of the failure of processing any miRNAs, although this could be linked to other factors, such as the tested cell type. An additional method validates miRNAs through an exogenous expression system [7]. To this end, the miRNA precursor sequence and its flanks are transfected via an expression vector into a cell culture and over-expressed. The detection of both the precursor and mature miRNAs via northern blotting is then used as requirement to validate the finding. The drawback of this method is that it validates miRNAs in an exogenous manner since northern blotting requires a high miRNA concentration. Validation in an endogenous manner is also feasible, but only for a limited set of strongly expressed miRNAs.

miRNA target validation As for miRNA themselves, miRNA-target interaction predictions are numerous, with over 150 million interaction sites reported in human [287]. It is however known that many of these interaction sites are false positives [288]. Therefore, experimental validation of such interactions is of utmost importance. A multitude of experimental methods have been employed to validate miRNA-target interactions and their results have been collected in two major databases, TarBase [289] and miRTarBase [290]. The latter contains in its latest version (8.0) 380,639 validated miRNA-target interactions in human. Since these validation methods have different advantages and drawbacks, miRTarbase has classified them into two categories, "strong evidence" and "less strong evidence". Among the "less strong evidence" methods are validation by microarrays or NGS, with NGS accounting for 359,298 (94.4%) of all interactions. Different NGS-based techniques have been developed to evaluate the miRNA-mRNA target landscape, mostly based on CLIP-seq (cross-linked immunoprecipitation followed by next generation sequencing) methods [291]. Most of these methods allow to detect miRNAs and mRNA targets that are bound to the RISC without identification of explicit pairings. Newer methods like CLASH [245] or CLEAR-CLIP [244] try to address this issue by additional linking of the miRNA to their mRNA target. Although these methods reveal thousands of miRNA-target interactions, their capture rate is still low, and they are affected by false positives. "Strong evidence" validation methods, which are all low throughput, are reporter assays, western blots, and RT-qPCR [292, 293]. They are used to show that changes in the miRNA expression affect the target mRNA/protein levels. Reporter assays are thus among the most reliable validation methods but limited in their throughput. This problem has been tackled recently by a target validation workflow enabling automated dual luciferase reporter assays [169].

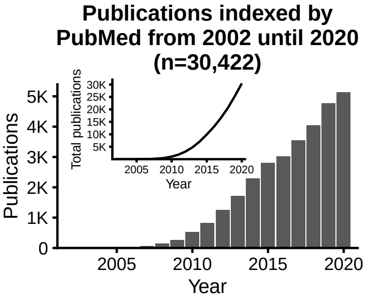


Figure 1.3: PubMed indexed publications containing the search term "miRNA" or "microRNA" in combination with the term "biomarker" shown per year as bar plot and, total number of articles shown as line plot.

Biomarkers A biomarker is "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease", as defined by the International Programme on Chemical Safety [294]. With the establishment of miRNA microarrays and subsequently NGS, the rising interest in miRNAs and the demonstration of their involvement in many diseases, the number of studies evaluating miRNAs as potential biomarkers increased exponentially, with a total of 30,422 publications at the end of 2020, as shown in Figure 1.3. Approximately 22% (30,422/138,648) of all miRNA related publications found in PubMed until the end of 2020 evaluate them in the context of biomarker research. In particular, miRNAs are considered promising non-invasive biomarkers, because they can be measured from various easily accessible body fluids, with the main focus lying on whole blood, plasma, and serum [170]. Saliva, urine, and seminal fluid have also been studied in this context [295–297]. The measurements of miRNAs circulating in blood, plasma, and serum are especially interesting since it has been shown that solid tissue-specific miRNAs can in part be detected in these fluids as well [9, 298]. Possible reasons include the increased permeability of the blood brain barrier in old diseased individuals [299] for brain specific miRNAs, but also the release of circulating tumor cells for tumor specific miRNAs [300] and the high vascularization of some tissues like the kidney or the liver [298]. Whole-blood biomarkers are promising as well since they additional convey information of the immune system. It is therefore not surprising that a plethora of publications evaluate miRNAs as biomarkers in the context of Alzheimer's disease [77, 301–304], Parkinson's disease [100, 101, 305–307], lung cancer [302, 308–312], chronic obstructive pulmonary disease [308, 313–316], breast cancer [317–321], cardiovascular diseases [322–326], and sports [327–331], among other. The use of miRNAs as biomarkers is also attractive because they have been shown to be stable under various conditions, such as different temperature, humidity levels [332], and repeated freeze-thaw cycles [333]. In addition, they are unaffected by ribonuclease degradation when circulating in extracellular vesicles or lipoproteins in various body fluids [334–336]. Nevertheless, the translation from research to clinical practice is still nearly non-existent, since only very few miRNA panels have been commercialized, which holds in particular for non-invasive miRNA panels [337, 338]. One reason is that many studies are performed on relatively small cohorts, often with small case numbers. Furthermore, different collection, storage, preparation, and analysis methods as well as different technologies limit knowledge transfer in meta-analyses [339]. This applies especially to studies performed with RT-qPCR, since those are limited to a small number of pre-selected miRNAs. Other factors include miRNAs that are generally associated with numerous diseases, for instance hsa-miR-144-5p [340] and hsa-miR-21-5p [341], which are thus not suitable as specific biomarkers. Such miRNAs are often found in studies with few patient groups, since their involvement in unspecific disease processes cannot be identified. Furthermore, some miRNAs have been shown to be highly correlated with age and sex, as well as the ethnic background [254, 342, 343]. Finally, most studies evaluate miRNAs in retrospective studies, while prospective studies are needed to obtain more accurate performance estimates.

Therapeutic targets The field of miRNA therapeutics mainly focuses on the reestablishment of expected miRNA expression levels through two aspects. One is the use of miRNA mimics to counter an observed down-regulation, while the other is the use of anti-miRs that down-regulate over-expressed miRNAs [344, 345]. This field is still in early development and no phase 3 clinical trial has yet been reached with any treatment [338, 346]. Challenges include the stability of the RNA molecules and their delivery, as well as potential side-effects caused by off-targets [338, 347, 348]. RNA molecules are typically unstable because of their 2'-OH group which enables RNA hydrolysis. This can be tackled by producing modified RNA molecules in which this group is replaced with a variant that prevents hydrolysis[349]. The delivery needs to be specific to the organ of interest and can be assisted, among others, with nanoparticles or specific molecules binding the targeted cells [350, 351]. One of the most advanced therapies is an anti-miR named Miravirsen, currently in phase 2, down-regulating miR-122 for treatment of hepatitis C virus infection [352].

1.2.4 *Profiling*

To investigate the role of sncRNAs, it is central to be able to measure and quantify them. Over the last decades multiple technologies have been introduced that enable their study at different levels, all with their own advantages and disadvantages. Typical steps, independent of the profiling technology, start with the collection of samples of the tissue, cell type or fluid of interest. Samples are then processed using RNA stabilizing solutions (using e.g., RNAlater for tissues and PaxGene blood tubes from Qiagen) and optionally frozen and preserved for further processing. Alternatives for blood are the collection of dried blood using filter paper [332], or with specific devices, such as the Mitra microsampling device [11]. This step is then followed by tissue and cell lysis, except in the case of blood, for which this step is typically combined with the RNA stabilization. Subsequently, an RNA extraction step is performed via phase separation, optionally combined with silica-based columns, glass-fiber filters, or a resin separation matrix [353]. After obtaining the total RNA of the samples, quality control steps are performed. This involves evaluating the RNA integrity, often using the Agilent Bioanalyzer, which performs an automatic electrophoresis, and provides length profiles of the evaluated RNA, as well an RNA integrity number (RIN), indicative of the progress of RNA degradation. In addition, RNA purity can be assessed with a spectrophotometer (e.g., Thermo Scientific NanoDrop). Further steps depend on the employed profiling technology. The most common profiling technologies encompass RT-qPCR, microarrays, and next-generation sequencing.

RT-qPCR The profiling of miRNAs by RT-qPCR is the oldest and best established of the common profiling methods. It can easily be performed in clinical laboratories since the needed instruments are broadly available, and employed for routine diagnostic tests. It allows to measure a small number of predetermined miRNAs. This technique consists of two separate steps. The reverse transcription of the RNA into complementary DNA (cDNA) with a reverse transcriptase is followed by the amplification of the cDNA through the polymerase chain reaction (PCR). PCR consists in a repeated number of thermal cycles triggering the denaturation of cDNA, annealing of complementary forward and reverse primers and elongation of the primers via a DNA polymerase that incorporates the added deoxynucleotide triphosphates (dNTP). The two most popular systems are the Thermo Fisher Scientific miRNA TaqMan assay and the Qiagen SYBR Green miSCript PCR system. The TaqMan assay employs a miRNA specific stem-loop primer during the reverse transcription phase and in turn uses miRNA specific forward primer and stem-loop specific reverse primers for the amplification phase. A miRNA specific TaqMan probe binds to the amplicons and emits a fluorescent signal when hydrolyzed by the polymerase, which can be used for quantification. The miSCript PCR system performs a polyadenylation of the miRNAs and uses a universal oligo-dT primer for the reverse transcription. Via a dual buffer system, either miRNA or mRNA can be converted into cDNA. For the PCR reaction a miRNA specific forward primer is used, as well as a universal reverse primer. The binding to double-stranded DNA of a fluorescent dye, SYBR Green I, emits light and is used for quantification. Two quantification methods are typically used [354]. The absolute copy number can be determined using a standard curve obtained through serial dilutions from known template amounts. This approach is less popular because it needs to be performed for each miRNA separately. The second method performs a relative quantification by comparing the quantification cycle C_q value reported for the RNA of interest to one or multiple C_q values of stably expressed RNAs. The C_q value is derived from the PCR amplification curve, and related to the number of cycles performed to obtain a signal stronger than a specific threshold. As calibrator RNAs either exogenous spike-in RNAs are used, such as miRNAs from other species that are not similar to any miRNA in the studies organism, or endogenous RNAs, often small nuclear RNAs (snRNAs) or snoRNAs, are used. Overall RT-qPCR has been shown to be highly sensitive and specific, while presenting a wide dynamic quantification range and good reproducibility [355]. Its disadvantage is the ability to only measure small miRNA panels. It is thus well suited for validation studies and clinical applications, but unsuitable for exploratory studies.

Microarrays The microarray technology became very popular for genotyping, gene expression profiling, and other genomics applications in the early 2000s [356–358]. Nowadays they are employed to detect or measure aspects of many additional omics types, such as proteins and peptides levels [359, 360], but also DNA methylation [361] or miRNA expression levels [362]. In the miRNA field it is used to profile large panels of miRNAs, typically all miRNAs present in miRBase of the species of interest. The basic principle for miRNA microarrays consists in the hybridization of labeled total RNA to miRNA specific probes on a chip and the measurement of the resulting light intensity. A chip consists of multiple arrays with at least one spot for each miRNA that is measured, and each spot is composed of a large amount of probes to which the sample RNA can bind. The two most popular microarray systems are the Agilent SurePrint miRNA microarrays and the Affymetrix GeneChip miRNA arrays [353]. The Agilent microarrays require an initial RNA dephosphorylation step at the 3′ end, followed by the ligation of a fluorescent dye, Cyanine 3 [362]. The probes consist of a complementary sequence to the miRNA, followed by a guanine that binds to the ligated cytosine, as well as an additional hairpin that improves miRNA binding specificity. After the ligation step, the hybridization is performed, which is then followed by washing all unbound RNA. Finally, the array is scanned with an Agilent Microarray Scanner and intensity values are measured for each spot [362]. The Affymetrix GeneChip miRNA array starts with a poly(A) tailing step followed by the ligation of a biotin-labeled dendrimer. Subsequently, the hybridization step is performed, followed by washing of unbound RNA and staining of the sample with Streptavidin, that will bind the biotin-labeled RNA. The array is then scanned with the GeneChip Scanner and expression intensities are measured [353]. To quantify the measured miRNAs, multiple steps are performed [363]. In a first step a background correction is applied. Then an aggregation step is performed, since typically multiple spots are measured for the same miRNA. This can be accomplished by averaging the background corrected signal or using the median polish method. In some cases, this is performed as the last step instead. Next a normalization method to enable the comparison of multiple arrays is performed. This is typically done via quantile normalization, but other methods like variance stabilizing normalization can also be employed. Finally, data are $\log_2$ transformed (except if they are already on a log-scale). An advantage of miRNA quantification by microarrays is high-throughput profiling of all known miRNAs at once, with up to 8 arrays per chip, depending on the manufacturer. In addition, no amplification step is necessary, thereby avoiding the introduction of biases related to this procedure. Agilent microarrays also showed in a study comparing them to RT-qPCR and NGS the highest reproducibility [355]. Disadvantages are that highly expressed miRNAs are less accurately quantified because they can reach signal saturation and low expressed miRNAs are limited by the background signal. Although custom microarrays can be designed to e.g., measure the expression of candidate miRNAs [4, 6], the *de novo* discovery of miRNAs with microarrays is not practical.

Next-generation sequencing Next-generation sequencing (NGS), also known as second generation sequencing, is a high-throughput technology originally developed to determine the DNA sequence of genomes. Since the commercialization of sequencers in the mid-2000s, the platforms have made major progress, especially in throughput and cost, going from over 3M dollars for a human genome in 2008 to less than 1,000 dollars in recent years (see Figure 1.4, [364]). While early sequencers generated less than one billion bases per run [365], they can now generate up to 3,000 billion bases [366]. Today a plethora of sequencing assays exist and cover a broad range of applications, such as, genotyping via whole-genome, or whole-exome sequencing, studying protein-DNA or protein-RNA interactions via chromatin immunoprecipitation sequencing, gene expression profiling via RNA-sequencing, or DNA methylation analysis via whole-genome bisulfite sequencing [367]. Importantly, assays to detect and measure sncRNAs are available as well. The most commonly used technology for these assays is commercialized by Illumina and is composed by three steps: (1) library preparation, (2) cluster generation by bridge amplification and (3) sequencing by synthesis [366]. First the 3′ and 5′ adapters are ligated to the RNA, then a reverse transcription of the RNA to cDNA is performed, followed by an amplification via PCR. During this amplification step, additional PCR index primers are added, specific for each library. This step is known as multiplexing and allows to identify each library after the sequencing is completed.

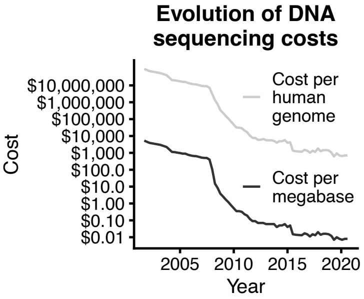


Figure 1.4: Plot showing the evolution of DNA sequencing costs, as collected by the National Human Genome Research Institute [364]. The cost per genome includes in addition to the number of needed megabases the sequence coverage needed for the assembly of a genome.

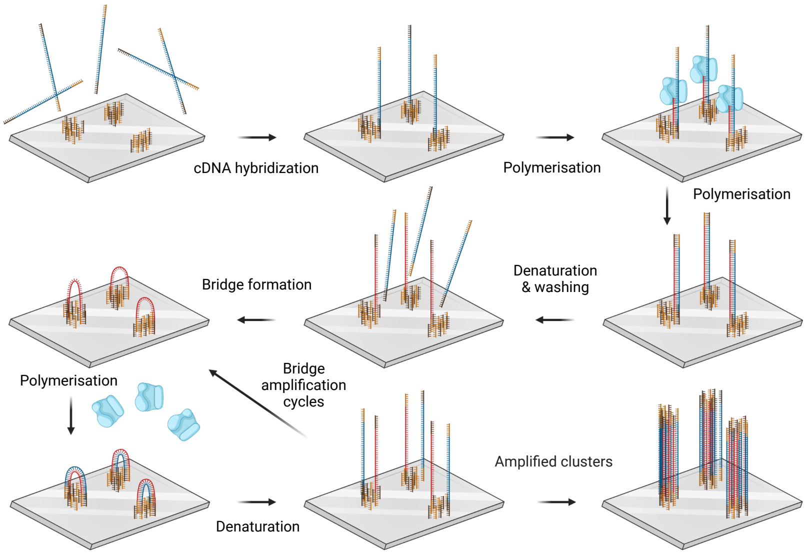


Figure 1.5: Scheme of the process of bridge amplification on a flow cell. First, the cDNA coupled with adapters hybridize to oligonucleotides tethered to the flow cell surface. The complementary strand is then polymerized, and the original strand is washed away after denaturation. The tethered ssDNA strand forms a bridge with another oligonucleotide of the cluster. After a polymerization step, the bridges are denatured, and the cycle can repeat itself, to form clusters of amplified DNA strands. Created with BioRender.com

Subsequently, a size selection of the cDNA is performed by cutting out the targeted fragment size after a gel electrophoresis or by performing a bead extraction. If multiple libraries are measured in parallel, they are typically pooled before or after the size selection step. Finally, the single or pooled library is normalized to ensure the same concentrations between all measured libraries. The library is loaded into a flow cell that will bind the cDNA sequences. Next, these sequences will be amplified via bridge amplification to generate a cluster for each sequence (see Figure 1.5). Before the sequencing step, all reverse strand products are cleaved and washed off and the 3′ ends of the bound sequences are blocked to prevent priming. Then follows the sequencing step, in which the sequencing primer hybridizes to the adapter sequence and a DNA polymerase adds fluorescently labeled dNTPs, one after the other. After the incorporation of one labeled dNTP, the fluorophore blocks further polymerization and emits a light signal that will be measured by the sequencer (known as base calling). Subsequently, the fluorophore is cleaved, and the next dNTP is incorporated. This step is repeated until the targeted sequencing length is reached (typically between 50 and 100 nucleotides for sncRNA sequencing). The sequences derived from this process are called reads. In case of paired-end sequencing, i.e., sequencing of the same sequence starting from the 5′ end, as well as from the 3′ end, a few more additional steps are performed. For sncRNA sequencing this is however rarely employed because the read length of single-end sequencing is sufficient to cover the length of most sncRNAs, especially when miRNAs or piRNAs are focused. Multiple quality criteria are subsequently evaluated after the sequencing run. The base calling accuracy is often calculated by adding spike-in sequences of the bacteriophage PhiX and determining the errors observed in the correspondingly measured reads. In addition, for each nucleotide of each read a Phred quality score Q is computed by the sequencer, which denotes the probability of a base-calling error P as

$Q = -10\log_{10}(P)$. Nowadays, a Q30 score, i.e., the percentage of bases that have a quality score of at least 30 and thus a base call accuracy of at least 99.9% is usually above 85% [368]. The quantification of sncRNAs is performed by either directly mapping the reads to specialized sncRNA databases (such as miRBase for miRNAs), or by mapping to the human genome and quantifying them accordingly in case they map to an annotated position. The obtained read counts are then usually normalized by the sequencing depth to reads per million (RPM) or by the number of reads mapping to the human genome (reads per million mapped, RPMM), or, when quantifying only specific sncRNAs, such as miRNAs, by the number of reads mapping to miRNAs (reads per million mapped to miRNAs, RPMMM). NGS presents multiple advantages over the other two presented technologies. It generates single nucleotide resolution, thereby allowing the quantification of isoforms and alternatively processed sequences. Moreover, it is not restricted by a finite set of sequences to be measured. Therefore, it can benefit from the most up-to-date annotations and custom annotations. It also provides the possibility to identify yet unannotated molecules [184, 192]. In addition, exogenous RNA can be identified, e.g., from bacteria or viruses, that might be related to the disease or represent a contamination [1]. Furthermore, similar to microarrays, it is possible to analyze multiple samples in parallel (the amount depending on the sequencing platform and library preparation method). In comparison to microarrays, NGS provides a wider dynamic range, as well as a higher sensitivity and slightly lower, but still high reproducibility [355]. This is also reflected by the availability of low-input library preparation protocols requiring only 50 pg RNA input for NGS [369], while microarray platforms typically need 100 ng RNA input [362]. However, drawbacks also exist. The technology is still more expensive than the more specialized RT-qPCR, it is time consuming, and automation is still in its beginning [370]. Also, it is affected by PCR amplification bias, which can lead to inflated counts of specific sequences [371]. This is currently being tackled by the introduction of unique molecular identifiers (UMIs) into the library sequences that enable a deduplication step after sequencing [372]. Moreover, the standard library preparation protocols do not allow the sequencing of 3′ phosphorylated RNAs, as well as 5′ capped miRNAs, and can be affected by ligation bias [16, 373]. Different protocols are currently evaluated to reduce these limitations [16, 374].

1.3 *Computational approaches*

The methods previously presented can generate large datasets with millions of data points that need efficient tools to be evaluated. In the following, an overview of the correct usage of tools for reproducible research and the most commonly applied analyses will be outlined, and the state of webserver and database development will be presented.

1.3.1 Reproducible research

In the last decade, the reproducibility of research results, or more specifically, the lack thereof, has gained increasing attention. In 2016, more than 70% of 1,576 of researchers across various fields stated that they failed to reproduce at least one experiment of another scientist [375]. The causing reasons are diverse and range from study design errors, the application of unsuitable methods and models, over wrong assumptions and invalid statistics, to data forgery [375–377]. Additional reasons stem from missing data or code availability, as well as a missing characterization of the software environment. Often, different versions of tools produce different results, which in turn might lead to other conclusions. Moreover, nowadays analyses become increasingly complex, and often combine tens or hundreds of tools and software packages. Therefore, an efficient organization and combination of analyses is highly important. This can be achieved through the use and implementation of standardized software pipelines, which is facilitated by workflow managers such as Snakemake [378], Nextflow [379], Cromwell [380] or Galaxy [381]. Those tools offer a framework to define, combine and execute processes. In particular, they provide the possibility to define and create independent environments for each process. To this end, various approaches are employed, such as software package management with Conda environments [382] or container-based virtualization technologies such as Docker [383] or Singularity [384]. This is also an advantage for the scalability of a workflow since the workflow manager can easily deploy the processes to multiple machines in a cluster or cloud instances without having to rely on a particular software environment. In addition to software versions, storing and sharing database versions is an equally important aspect. For instance, the gene and transcript annotations for the human genome can be retrieved from RefSeq [385], Ensembl [386], or GENCODE [387]. Each source frequently releases new and modified annotations, thus rendering reproducible analyses impossible when the source and release number are unknown [388, 389]. To this end, tools have been implemented that automatically store this information together with the generated results [390]. Additional steps that can be taken to ensure higher reproducibility include code versioning with version control systems such as Git and uploading bundled workflows, including all necessary input, code, and generated output files, to platforms storing research data, such as Zenodo or figshare.

1.3.2 sncRNA analysis

The typical analysis workflow for sncRNAs, with a focus on miRNAs, is presented in Figure 1.6. As shown, some analyses are reserved to NGS datasets, while others are shared by all profiling methods. These analyses are often implemented in analysis toolboxes such as sRNAbench [391], miRge [392] or miRMaster [1, 2]. In the following, each step will be briefly explained. First, for NGS datasets, the reads are demultiplexed to generate one FASTQ file per library, containing the sequenced reads and their base quality. Next, a quality control step is performed verifying the number of reads per library, the base quality distribution, the distribution of unidentified bases (labeled with N), the adapter content and the sources of over-represented sequences (possibly affected by adapters or primers). Popular tools supporting this step are FastQC and fastp [393]. Since the sequenced RNA molecules are often smaller than the read length, 3′ adapters can overlap into the read. Therefore, an adapter trimming step is required, which is often combined with a quality filtering step, to remove bases or reads with low quality. This is implemented in tools such as Cutadapt [394] and Trimmomatic [395]. After excluding reads shorter than 17 or 18 nucleotides, the quality control step is repeated to ensure that an adequate number of high-quality reads remains. Next, a quantification procedure followed by a normalization step is performed, as described in 1.2.4. For NGS data, after mapping the reads, usually with the tools Bowtie [396], Bowtie 2 [397], BWA [398],or STAR [399], an additional quality control step is performed, evaluating the mapping statistics and the percentage of reads that could be attributed to a known RNA class. Furthermore, to reduce data noise in the subsequent analyses, sncRNAs that are expressed below a determined detection threshold in most samples are removed. Once all profiled RNAs have been quantified, various analyses are possible.

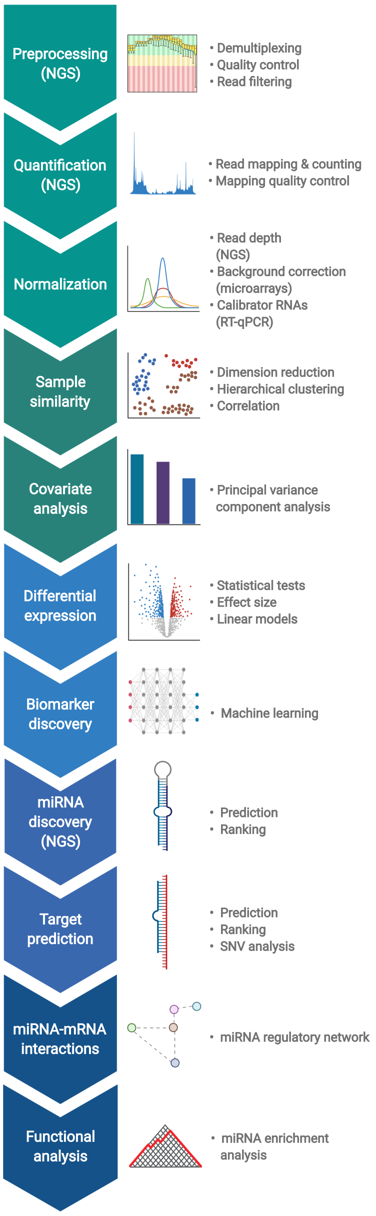


Figure 1.6: Schematic representation of the steps of a typical sncRNA analysis workflow. The major steps are outlined on the left. Technology-specific steps or routines are labeled correspondingly. Each step is composed of one or more possible routines. Created with BioRender.com

Sample similarity First, because of the high data dimensionality (thousands of dimensions for microarray and NGS data), a visual representation of the distance or similarity of the expression profiles between the measured samples is usually achieved through the application of unsupervised dimension reduction and clustering methods. The most commonly applied dimension reduction method is principal component analysis (PCA). Briefly, this method computes a linear transformation of the data that ensures that each resulting feature (component) captures most variance, while being orthogonal to all others. Therefore, often only the first two components, which explain the largest fraction of observed variance, are plotted. Since PCA performs a linear transformation, it fails to uncover non-linear relationships. To this end, other data reduction techniques were developed, such as k-neighbor based graph learning algorithms like t-distributed stochastic neighbor embedding (t-SNE) [400] and uniform manifold approximation and projection (UMAP) [401]. t-SNE first computes a neighborhood-aware probability distribution derived from the similarity between each pair of samples. It then creates a random low-dimensional representation as starting point and computes the probability distribution for it. Next, via gradient descent, it minimizes the Kullback-Leibler divergence which measures the difference between the probability distributions of the low-dimensional embedding and the original data. As a result, t-SNE produces low-dimensional representations that preserve local relationships. UMAP is a dimension reduction method that also produces embeddings preserving local relationships. Moreover, global relationships can be represented by UMAP as well, which is why it is often preferred to t-SNE, although this is strongly influenced by the parameters used by the algorithm [402, 403]. Another advantage of UMAP over t-SNE is that it can be computed faster and requires less memory [401]. Another approach to determine the similarity of expression profiles is to perform hierarchical clustering of the expression profiles (potentially restricted to the most varying sncRNAs). A third approach is to cluster the correlation coefficients of the expression profiles. The Pearson's correlation coefficient is computed when linear relationships are expected, or the data is log-transformed. Alternatively, Spearman's correlation coefficient is computed, which corresponds to the Pearson's correlation coefficient on the ranks of the variables, thereby allowing it to capture non-linear similarities.

Covariate analysis Biological variables such as the tissue of origin, the characterized cell type, the age, or sex can influence the observed expression patterns. But technical batches can also contribute to observed differences. Therefore, it is important to estimate the contributions of each known factor to the measured profiles. To this end, principal variance component analysis can be performed, which allows to estimate the proportion of variance explained by each factor of interest [404]. This technique combines principal component analysis and variance component analysis. Briefly, first principal components are obtained via PCA. Then, mixed linear models including the variables of interest as random factors, as well as two-way interactions between them, are fitted to the principal components. Subsequently, the variance of each factor for each principal component is extracted from the model and standardized. Finally, the average variance proportion weighted by the proportion of explained variance of each principal component is computed for each factor.

Differential expression Differential expression analysis is performed through statistical analysis and allows to determine which sncRNAs exhibit different expression levels between the tested experimental conditions. To this end, various statistical tests are commonly employed. When the data is normally distributed, the Student's *t*-test can be employed to determine the significance of the difference observed between two experimental conditions, while a one-way analysis of variance (ANOVA) can be applied when two or more conditions are compared. In case the data does not follow a normal distribution, which can be tested by the Shapiro-Wilk test, non-parametric tests such as the Wilcoxon-Mann-Whitney test can be used to compare two conditions, or in case of more than two conditions, the Kruskal-Wallis test can be employed. Statistical significance can also be computed through generalized linear models, as implemented in the R packages Limma [405], edgeR [406], and DESeq2 [407]. Such models can be particularly useful when the effects of some covariates are large and need to be considered explicitly. Finally, since the likelihood of obtaining a significant P-value by chance rises with the number of tests performed, a P-value adjustment method that corrects for such errors is required.

The most commonly applied method is the Benjamini-Hochberg procedure which corrects for the false discovery rate [408]. Complementary to these tests, fold changes and effect sizes such as Cohen's D or the area under the curve (AUC) should be computed, to assess the size of the observed differences [409]. This is particularly important, since P-values depend on the cohort size and thus arbitrarily low P-values can be reached, albeit only small differences are present.

1.3.3 *RNA biomarker discovery*

Biomarker discovery is typically first performed on measurements taken via microarray or NGS assays to enable a broad selection of candidates. Validation studies usually build on signatures found in exploratory studies and measure these via RT-qPCR. While biomarkers can be identified independently via differential expression, single sncRNAs are often not powerful enough to reliably identify a biological condition. Therefore, the identification of a signature of multiple sncRNAs is a central aspect of biomarker discovery. To this end, machine learning algorithms for binary classification are applied. These algorithms derive models from the provided input data that classify the data into two classes, typically denoted as positive and negative class. For instance, for the discovery of diagnostic biomarkers, the group of participants with the disease of interest is assigned to the positive class, while the comparison group is assigned to the negative class. The applied algorithms range from linear models such as logistic regression [410], linear support vector machines [411, 412], and elastic nets [413, 414], to algorithms capturing non-linear effects, such as support vector machines with a Gaussian kernel [411, 412], random forests [412, 415], gradient boosted trees [25, 416], and (deep) neural networks [417].

To identify signatures of reasonable size that perform well, a feature selection procedure is applied. This procedure either consists of filter methods, wrapper methods, or embedded methods [418]. Filter methods are usually fast to compute and assign a score to each feature that measure its usefulness. Common techniques include statistical tests, mutual information, and Relief-based algorithms [419], which take feature interactions into account. Wrapper methods are computationally expensive since they iteratively evaluate feature subsets with a machine learning model. Commonly applied techniques often use greedy approaches or optimization algorithms to soften the computational impact, since an exhaustive evaluation of all possible subsets is quickly unfeasible, due to the exponential growth of possibilities. An example of such a technique is forward selection, which starts from an empty model and selects one additional feature at a time, after having determined which added feature results in the best model performance. Therefore, for a dataset with p features, the maximum number of fitted models is $\sum_{k=0}^{p}(p-k) = 1 + \frac{p(p+1)}{2}$. Thus, when selecting features among e.g., 1,000 miRNAs, "only" 500,501 models need to be evaluated, instead of $2^p = 1.07 \cdot 10^{301}$ models. The addition of features is typically stopped when the model is not improving anymore. Another common wrapper-based feature selection algorithm is Recursive feature elimination (RFE) [420]. This algorithm starts with a model trained on all features and removes at each step the feature determined to be the least important by the model. The advantage of this algorithm is that at most p models need to be fitted for p features. The final category of feature selection algorithms, embedded methods, represents a fast alternative to wrappers, but restricts the choice of classifiers. Typically, such models use a regularization mechanism to shrink coefficients of unimportant features to zero, thereby omitting features not contributing to the model. Alternatively, tree-based models, such as random forests, or gradient boosted trees provide feature importance measures based on the trained model and thus allow to select a feature subset after fitting the model only once.

To evaluate the performance of a machine learning model the cohort is typically split into two sets, a training set and a validation set. The first one is used to train the appropriate model, while the second one is used to evaluate the performance of the model. To increase the reliability of the performance estimates resampling methods like k-fold cross-validation can be applied. In this case, the dataset is divided into k subsets and the performance evaluation process is repeated k times, each time with another subset being used as validation set, while the remaining ones are used to train the model (as illustrated by Figure 1.7). The estimates of k-fold cross-validation can be further improved by repeating the process multiple times, leading to different splits for each k-fold cross-validation evaluation.

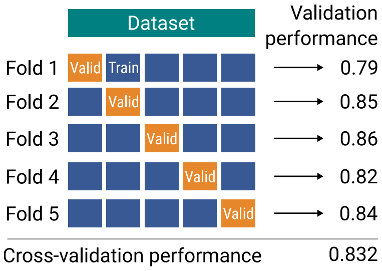


Figure 1.7: Schematic representation of five-fold cross-validation. For each fold, the performance metric is computed on the validation set and the cross-validation performance is reported as the mean over all folds.

For an unbalanced cohort with e.g., 10 times more healthy patients than diseased ones, certain precautions need to be taken. In such classification scenarios, machine learning models are prone to simply predict only the majority class and thus lead to weak models. This can be tackled with multiple methods [421]. One consists in artificially altering the ratio of the considered classes by over or under-sampling. Alternatively, the misclassification cost can be modified in favor of the underrepresented class. In addition, tuning the class probability cutoff returned by the model can lead to a balancing of the assignments to the underrepresented class.

Finally, the model performance needs to be measured with adequate metrics. For balanced datasets, and in case of equally important classes, the accuracy, defined as the ratio of correct predictions among all predictions is a commonly used metric. However, datasets are often unbalanced, and the detection of one class is more important than the other. In this case, the area under the curve of the receiver operating characteristic curve (AUC-ROC) is a more suitable measure. The AUC-ROC takes values between 0 and 1, where 0.5 describes the performance of a random classifier. Other measures to evaluate are the sensitivity (i.e., the fraction of true positive predictions among all positive cases) and the specificity (i.e., the fraction of true negative predictions among all negative cases). While the computed performance metrics are an indicator of the generalization performance of a classifier, they are tied to the prevalence of the group of interest. Depending on the cohort design, the prevalence in the cohort might not reflect the prevalence in the real diagnostic scenario. Therefore, additional measures that take into account the real prevalence, such as the positive and negative predictive value, can help to estimate the performance of the classifier under different diagnostic scenarios.

1.3.4 Novel miRNA discovery

The discovery of uncharacterized miRNAs is a central aspect of miRNA research, especially important for poorly characterized organisms, but also relevant to well studies species such as human and mouse. An initial *in silico* prediction is typically performed before time-consuming and expensive experimental validation methods. The identification of novel miRNAs is hampered by their small size, but also by the fact that they can be located at nearly any position in the human genome. This challenge is reflected by the predictions of thousands of potential miRNAs in many studies, of which many reveal to be false positives [7, 184, 191]. To improve the identification of miRNA candidates, many prediction tools start with the identification of pre-miRNAs. Therefore, those tools can take advantage of known properties of pre-miRNAs, such as the formation of a characteristic hairpin structure, low minimum free energy [422], and the presence of sequence motifs in their vicinity [423–425]. Overall, miRNA prediction tools can be grouped into three categories. The first category is composed of homology-based tools relying on the evolutionary conservation of miRNA sequence and structure. Among those are the early prediction tools, miRScan [426], and miRSeeker [427], both published in 2003, which were used to predict miRNAs in *C. elegans* and *Drosophila* species, respectively. One drawback of these methods is that they cannot identify species specific miRNAs. The second category of tools is composed of *ab initio* prediction tools, mostly employing machine learning models that use sequence and structure features of precursor miRNAs. These tools commonly use support vector machines or random forest models [428–430] but have most recently also applied deep learning methods such as convolutional neural networks and recurrent neural networks [431, 432]. The third category is composed of tools taking advantage of NGS data. In particular, this enables the incorporation of knowledge of the biogenesis of miRNAs. As a result of the processing of the pre-miRNA with Dicer, reads of appropriate length mapping to real mature miRNA are expected to accumulate on the 5′ and 3′ arms of the hairpins, while only few reads are expected to map across or in between these regions. Furthermore, knowledge about isomiRs can be integrated. Therefore, reads are expected to begin at few different positions at the 5′ end of the miRNA candidates, as opposed to the 3′ ends. The most popular tools are miRDeep2 [285], sRNAbench [391] and miRge [392]. The disadvantage of these tools is that they require an assembled genome. Therefore, reference-free miRNA prediction methods were developed, such as miReader [433], MirPlex [434] and Mirnovo [435]. Since pre-miRNAs usually cannot be identified from

NGS reads only, and thus no secondary hairpin structure can be computed, these tools focus on miRNA duplexes or read profile patterns. Because of missing additional information, these tools usually perform more poorly [435]. Yet, the assessment of the performance of miRNA prediction algorithms is highly dependent on the validation sets used. Datasets leading to the most accurate false positive estimates would be composed of miRNA-like sequences that could not be validated experimentally. However, such datasets are rare. Therefore, the performance of algorithms is often only measured according to the obtained sensitivity, or by evaluating pseudo pre-miRNA sets derived from hairpin like sequences which are not expected to produce any miRNAs [1]. Although prediction algorithms have improved over the last decade, they still lead to a large number of false positives. To this end, approaches that score and filter miRNA candidates are available [184].

1.3.5 miRNA target prediction

One of the first steps to determine potential biological functions of miRNAs is the identification of their target genes. Although human miRNAs are among the best characterized miRNAs, the function of many is still unknown. This is reflected by only 735 out of 2656 miRNAs for which at least one target was experimentally validated with strong evidence in the latest release of miRTarBase, as well as functional Gene Ontology term associations existing currently for only 400 miRNAs. However, as opposed to plants, where miRNAs bind with nearly perfect complementarity [436], miRNA targeting is highly flexible in mammals, leading to thousands of potential binding sites. Therefore, computational methods are necessary as first filter step, before extensive experimental validation can be performed. Target prediction tools typically incorporate at least one of four binding site targeting properties: (1) sequence complementarity, (2) formed secondary structure and thermodynamic stability, (3) accessibility, (4) binding site and miRNA seed conservation [437, 438]. To date, over 100 prediction tools have been implemented, most reporting only moderately overlapping binding sites or targets [287, 437]. Therefore, consensus approaches are often used as method to reduce the number of false positive interactions. Nevertheless, some tools are more commonly used since they appear to perform well in certain scenarios. Such examples are miRanda [439] and TargetScan [240]. miRanda uses a dynamic programming approach to find binding sites, based on sequence complementarity and secondary structure properties. TargetScan was originally published in 2003 [239] and continuously improved, with version 7 published in 2015 [240]. It incorporates all binding site targeting properties into a scoring-based algorithm. Overall, like the assessment of miRNA prediction algorithms, the assessment of target prediction algorithms is difficult, since negative datasets, i.e., datasets with potential target that contain non-functional binding sites are rarely published [440]. Therefore, prediction algorithms are often evaluated in combination with over-expression experiments (typically measured with microarrays), in which the fold-change induced by the over-expression of a particular miRNA is measured and used as proxy for the likelihood of correctly predicted target genes [240]. Since miRNAs are expected to mostly down-regulate mRNAs, high-fold changes are thought to be the result of the regulatory effect of the over-expressed miRNA. As a result of the high false positive rate of the currently available algorithms, ranking tools have been developed to efficiently prioritize validation experiments [441, 442]. In addition, a recent study demonstrated that filtering targets for genes involved in the same pathway can further raise the validation success rate [169].

1.3.6 SNV-induced miRNA targetome

With currently over 500 million known single nucleotide variants (SNVs) in the human genome [278], it is evident that a subset might affect miRNAs themselves or their binding sites. To evaluate the impact of such variants, the targeting tools introduced in the previous subsection are typically evaluated on the alternative sequences [21]. The difference in the predicted binding sites leads to the derived estimates of gained or lost targets. Therefore, these evaluations are tightly coupled with the quality of the predictions of the employed tools. Because of the possible co-occurrence of multiple variants, the number of potential alternative sequences grows exponentially. Therefore, most tools and databases evaluate only the impact of single variants in binding sites or miRNAs separately [443, 444]. This research field is strongly dependent on the target prediction field and thus runtime and performance improvements of target prediction tools will enable better and possibly more complete estimates of the impact of SNVs.

1.3.7 miRNA regulatory network modeling

Since miRNAs orchestrate gene regulation together, with several miRNAs potentially regulating the same target, their regulatory landscape is highly complex. To gain an understanding of central regulatory elements and their potential function, this landscape is typically modelled via regulatory networks. These networks often take the form of bipartite graphs, i.e., graphs allowing only connections between two distinct set of nodes, which are on one hand the miRNAs of interest, and on the other hand their target genes [445–447]. Tools that implement such approaches include MAGIA [448] and mirConnX [449]. Regulatory networks can grow very large, when connections are only based on predicted or validated targets. However, neither all target genes, nor all miRNAs are always expressed. Moreover, predicted targets or target genes that were validated *in vitro* do not necessarily interact with the corresponding miRNA *in vivo*. Therefore, expression data for at least miRNAs or target genes, and in the best case for both, allow to greatly reduce false positive connections in the graph. Additional filtering can be performed by only keeping interactions that reflect a potential down-regulation, because of an observed negative correlation between the miRNA and target gene expression values. The obtained network can then be examined for sub-modules and regulatory motifs via graph theory [450]. The thereby obtained sub-networks can be further analyzed for function via enrichment analyses [445].

1.3.8 miRNA enrichment analysis

After the identification of miRNAs of interest, determining their functional involvement plays a central role in many study setups. Enrichment analysis is a method that finds statistically significant over-representations between the miRNAs of interest and known miRNA or gene categories, representing for example biological processes and pathways. It is typically implemented either via an over-representation test such as the one-sided Fisher's exact test or via Gene Set Enrichment Analysis (GSEA) [451]. While the over-representation analysis is performed on selected subsets (e.g., all significantly deregulated RNAs), GSEA is performed on the list of all measured RNAs, ranked by a criterion such as the fold change. It reports for each of the evaluated categories if the RNAs of the corresponding category accumulate at the beginning or end of the ranked list. The significance of such accumulations can be estimated via permutation tests, or in certain cases, can be computed exactly with a dynamic programming approach [452]. Enrichment analysis is well known from transcriptome analysis and has been implemented in multiple tools such as GSEA [451], PANTHER [453], WebGestalt [454] and GeneTrail [455]. Among the most popular knowledge bases included by these tools are the Gene Ontology (GO) [456], containing the largest number of hierarchically organized annotations of gene functions and gene products, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [457] and Reactome [458], both collecting biochemical and disease-related pathways. These tools are however not directly applicable to miRNA research. In particular, early approaches that performed indirect enrichment by simply considering all target genes of the miRNAs of interest raised concerns because of spurious categories being reported even for random sets [459, 460]. With the development of databases directly annotating miRNAs, such as miR2Disease [461] and the Human MicroRNA Disease Database (HMDD) [462], both collecting sets of miRNAs playing roles in diverse diseases, and the development of computational approaches to handle the observed biases [459], the results produced by miRNA enrichment tools became more reliable. Examples of tools implementing appropriate approaches are TAM [463] and miEAA [22, 183]. Overall, miRNA enrichment tools are gaining importance, especially with an increasing number of resources collecting directly related information, and the continuously growing number of validated targets. This is also reflected in the increasing popularity of databases hosting miRNA pathway information, such as DIANA-miRPath [464] and miRPathDB [19, 20].

1.3.9 Webserver and database development

While the development and creation of new software packages, command line tools, and datasets are of utmost importance, their usage is often limited to scientists with programming knowledge. To make novel and state-of-the art software or datasets available to a broader community, graphical user interfaces are needed, and as few computer knowledge as possible should be required. The most user-friendly way to provide access to new software or datasets is via webservers. This requires no special hardware, only a web browser, and allows to skip any tedious installation procedure. It also leaves all the dependency management and other requirements to the developers, which can test their server in a well-defined environment. Webservers and databases are highly relevant in life sciences and biomedical research. Among the most prominent are databases hosting genome annotation information such as RefSeq [385] or Ensembl [386], databases hosting protein sequences and structures such as UniProt [465] and the Protein Data Bank (PDB) [466], but also portals providing access to a large range of experimental data such as The Cancer Genome Atlas (TCGA) [467], the Encyclopedia of DNA Elements (ENCODE) [468], and the Genotype-Tissue Expression (GTEx) portal [469], but also the Gene Expression Omnibus (GEO) [470], and the Sequence Read Archive (SRA) [195]. Widely used webservers include the sequence analysis tool BLAST [471], the phylogenetic analysis software PhyML [472], and the protein similarity search tool HMMER [473]. As a result of the technological advances of the last decade, webserver development has evolved in many aspects. While early webpages presented mostly static content, nowadays interactive representations and client-side processing are at the forefront, backed by the rapid evolution of the browser scripting language JavaScript and the continuous improvements of Cascading Style Sheets (CSS)). This evolution has fostered the development of widely used libraries such as jQuery (used by 78.1% of all websites and 95.7% of all websites using JavaScript [474]), which provides basic functionality to manipulate the HyperText Markup Language (HTML) Document Object Model (DOM), but also facilitates event handling, CSS animations and Ajax web development. In addition, large frontend styling frameworks have been developed, such as Bootstrap, which can be found on 22.2% of all websites [474]. Furthermore, with the growing popularity of JavaScript (ranking as the most popular language on GitHub since 2014, according to the number of repositories created), large frontend frameworks such as Angular (backed by Google) and React (backed by Facebook) were created. Similarly, many backend frameworks in popular programming languages were created and updated to accommodate the evolving requirements, among which the most popular (according to the number of stars on GitHub) are Laravel (PHP), Flask and Django (Python), ExpressJS (Node.js), and Ruby on Rails (Ruby). Moreover, webserver development has also benefited from container virtualization, allowing to easily run multiple webservers with different environments on the same machine. This has also been reflected in the integration of advanced DevOps into software development platforms such as GitHub and GitLab, which enable continuous integration (i.e., testing) and deployments of webservers. With the rising popularity of webservers accompanying scientific manuscripts, programmers with no background in web development are often required to create interactive webpages. To this end, the software packages Dash and Shiny in the popular data analysis scripting languages Python and R were created. Those packages hide most of the frontend to backend communication and provide components to easily model common visualization scenarios.

2
Goals of the PhD thesis

The major goal of this thesis was to improve our knowledge about miRNA characteristics and functions, as well as their role in clinical research. To reach this goal, five complementary objectives were pursued. In the following, each objective and the associated publications will be presented. A chronological overview is given in Figure 2.1.

The first objective was to create computational software allowing to comprehensively evaluate miRNA high-throughput sequencing data. To this end, a webserver that could perform most of the analyses presented in 1.3.2 was created in 2017 (miRMaster, [1]) and updated in 2021 (miRMaster 2, [2]). Furthermore, a webserver to efficiently analyze miRNA arm shifts based on the output of tools like miRMaster was implemented (miRSwitch, [3]).

The second objective was to enhance our understanding of the elements that define a miRNA. Thus, a large collection of human NGS sequencing data was evaluated with miRMaster, which resulted in a study evaluating the distribution and patterns of sncRNAs in the human genome, and predicting a large collection of miRNA candidates [4]. This led to a database, miRCarta, incorporating the read patterns and the discovered miRNA candidates [5]. We further investigated the role of a subset of these miRNA candidates in various diseases [6]. These studies resulted in a large validation study, which allowed us to provide an estimate of the number of human miRNAs [7]. In parallel, we investigated the expression and distribution of miRNAs in a large variety of human tissues [8], as well as body fluids [9], and evaluated the influence of seasonal changes on miRNA expression profiles [10]. Moreover, we evaluated sncRNA expression and conservation patterns in the blood of zoo animals [11], and generated an expression atlas of sncRNA profiles in mouse tissues [12].

The third objective was the assessment of technological advances and technical effects on the profiling on sncRNAs. Therefore, a new amplification-free sequencing technology published by BGI was evaluated [13]. Furthermore, the effects of RNA integrity on sncRNA data analysis were inspected [14]. In addition, a new low-input library preparation protocol was investigated in the context of dried blood sampling [15], and compared to other protocols [16]. Finally, a new sequencing chemistry based on antibody labeling was examined [17].

The fourth objective was the development of tools and resources to study the orchestrated targeting of genes and pathways performed by miRNAs. For this purpose, a webserver modelling miRNA interaction networks was implemented (miRTargetLink, [18]), and a database collecting miRNAs and their targeted pathways was created in 2016 (miRPathDB, [19]) and subsequently updated in 2019 (miRPathDB 2, [20]). Furthermore, a database collecting the information of the influence of SNVs on the human miRNA targetome was built (miRSNPdb, [21]), and a new version of the miRNA set enrichment analysis tool miEAA was published (miEAA 2, [22]).

The fifth and last objective was to determine the suitability of miRNAs as biomarkers for human diseases. To this end, blood miRNA profiles of large disease cohorts were investigated with machine learning methods. In particular, miRNA profiles were evaluated for neurodegenerative diseases like Alzheimer's disease [23], and Parkinson's disease [24], but also for lung cancer patients [25]. Finally, the influence of aging on miRNA profiles was investigated [26].

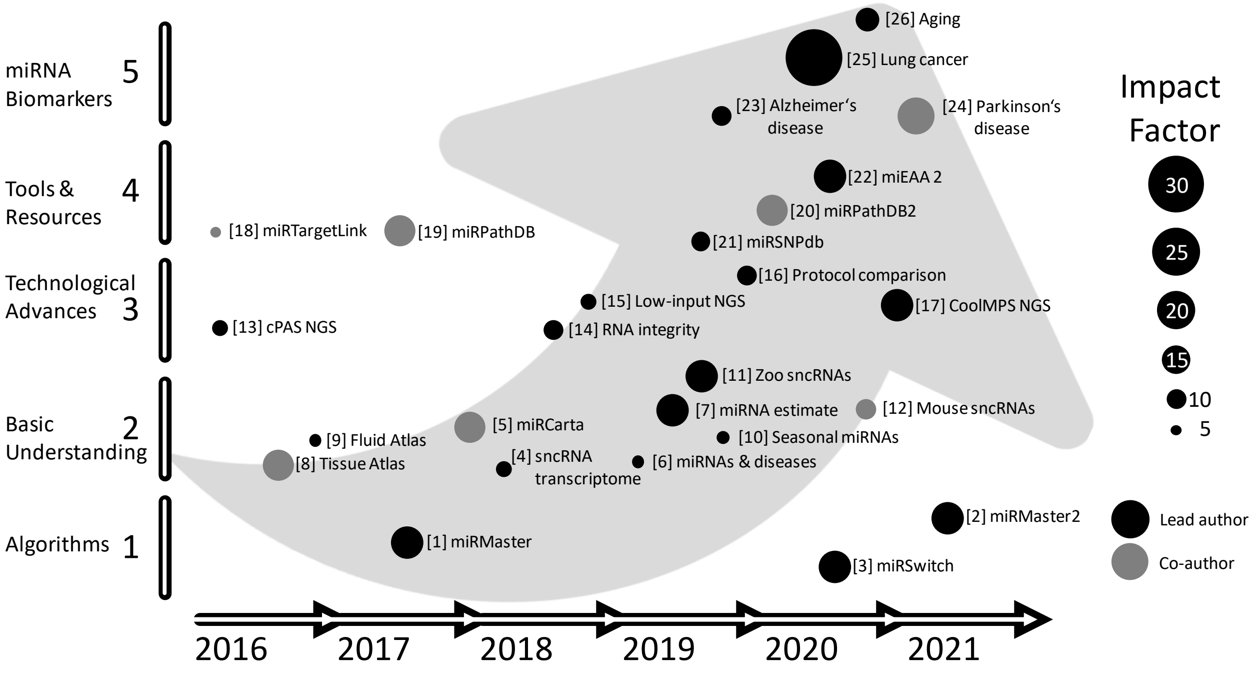


Figure 2.1: This plot shows the chronological order of the publications covered in this thesis, grouped by objective. The dots are colored by author role and the size of the dots shows the impact factor (IF) of the journals the articles were published in. Since no IF was available for Nature Aging, it was interpolated based on the citation counts of the articles published in the first issue.

In addition to the publications presented in this thesis, other contributions were made to related fields. Those span studies investigating miRNA targeting [169, 246, 437, 475–478] and miRNAs in various diseases and sports [313, 317, 329, 332, 479–484]. Moreover, contributions were made to studies evaluating technical effects on miRNA discovery [485], characterizing the expression of miRNAs in blood cells [486], as well as in the organs of the human mummy Ötzi [487]. Further studies include the evaluation of miRNAs in the context of other omic types [488, 489] and reviews of miRNA analysis tools [490, 491]. Technological advances such as single-cell miRNA sequencing methods were investigated as well [492]. The state of webservers in biology was reviewed, and a tool to monitor these webservers was published [493, 494]. Moreover, support was provided for studies investigating deep learning methods in the clinic [495, 496] and decentralized privacy-aware machine learning [497]. Additional studies tackled genomic variants in cardiomyopathy [498], copy number detection in stem cells [499], as well as tools and resources for bacterial research [500, 501]. Contributions were also made to a statistical tool evaluating the significance of ordered set overlaps [502]. A new technology for single-nuclear RNA sequencing was evaluated [368], as well as the effect of aging on the gene expression in mouse organs [111]. Furthermore, the development of a scientometric analysis tool [503], as well as scientometric analyses of research in atrial fibrillation and coronavirus disease 2019 (COVID-19) were supported [504, 505]. Finally, contributions were made to a study investigating the influence of COVID-19 on the gene expression in brain cells [506].

3 Results

This cumulative thesis is based on 26 peer-reviewed publications whose published versions are included in this chapter.

Published online 12 July 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 15* ***8731–8744***
doi: 10.1093/nar/gkx595

Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs

Tobias Fehlmann[1,*], Christina Backes[1], Mustafa Kahraman[1,2], Jan Haas[3,4,5], Nicole Ludwig[6], Andreas E. Posch[7], Maximilian L. Würstle[8], Matthias Hübenthal[9], Andre Franke[9], Benjamin Meder[3,4,5], Eckart Meese[6] and Andreas Keller[1]

[1]Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany, [2]Hummingbird Diagnostics GmbH, Heidelberg, Germany, [3]Department of Internal Medicine III, University Hospital Heidelberg, Heidelberg, Germany, [4]German Center for Cardiovascular Research (DZHK), Heidelberg, Germany, [5]Klaus Tschira Institute for Integrative Computational Cardiology, Heidelberg, Germany, [6]Department of Human Genetics, Saarland University, Homburg, Germany, [7]Ares Genetics GmbH, Vienna, Austria, [8]Siemens Healthcare GmbH, Strategy and Innovation, Erlangen, Germany and [9]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany

Received April 25, 2017; Revised June 23, 2017; Editorial Decision June 26, 2017; Accepted July 04, 2017

ABSTRACT

The analysis of small RNA NGS data together with the discovery of new small RNAs is among the foremost challenges in life science. For the analysis of raw high-throughput sequencing data we implemented the fast, accurate and comprehensive web-based tool miRMaster. Our toolbox provides a wide range of modules for quantification of miRNAs and other non-coding RNAs, discovering new miRNAs, isomiRs, mutations, exogenous RNAs and motifs. Use-cases comprising hundreds of samples are processed in less than 5 h with an accuracy of 99.4%. An integrative analysis of small RNAs from 1836 data sets (20 billion reads) indicated that context-specific miRNAs (e.g. miRNAs present only in one or few different tissues / cell types) still remain to be discovered while broadly expressed miRNAs appear to be largely known. In total, our analysis of known and novel miRNAs indicated nearly 22 000 candidates of precursors with one or two mature forms. Based on these, we designed a custom microarray comprising 11 872 potential mature miRNAs to assess the quality of our prediction. MiRMaster is a convenient-to-use tool for the comprehensive and fast analysis of miRNA NGS data. In addition, our predicted miRNA candidates provided as custom array will allow researchers to perform in depth validation of candidates interesting to them.

INTRODUCTION

MicroRNAs (miRNAs) play a central role in orchestrating human gene regulation and are consequently prime targets in biomedical research. Many miRNAs from *Homo sapiens* and other species are collected in the miRBase (1). Currently, the fraction of actually true positive miRNAs in this database is controversially discussed (2–10), especially later versions seem to contain many false positives (11). On the one hand, this calls for curated databases, on the other hand not all miRNAs, especially context specific ones, seem to be discovered yet.

Various experimental approaches are applied for measuring miRNA expression levels including approaches for small sets of selected miRNAs like RT-qPCR, CMOS based assays (12) or immunoassays (13). The most frequently employed genome-wide assays include microarray screening and high-throughput sequencing (HT-seq). A comparison of 12 different experimental approaches is provided by Mestdagh *et al.* (14).

HT-seq enables—beyond quantitative analysis of known miRNAs—single-base resolution of known and novel miRNAs (15) and thus is currently applied to discover the afore mentioned context-specific miRNAs. For the analysis of HT-seq data, a wide range of stand-alone and web-based bioinformatics tools have been implemented allowing the prediction of novel miRNA candidates and quantification of miRNAs (16,17), detection of miRNA isoforms (18,19), miRNA set enrichment analyses (20,21), and prediction of miRNA targets (22,23) among others. Akthar *et al.* published a comprehensive review on 129 available miRNA bioinformatics tools (24). The different data formats used in these tools and the challenges to combine web-based and stand-alone solutions, however, complicate the design of integrated pipelines.

[*]To whom correspondence should be addressed. Tel: +49 681 30268603; Email: tobias.fehlmann@ccb.uni-saarland.de

Our ambition was to develop a web-based application that combines the most frequently requested analyses. An important aspect of our tool termed miRMaster (www.ccb.uni-saarland.de/mirmaster) was to facilitate HT-seq data analysis of human samples from raw sequencing files provided in the FASTQ format. Building up on the basic principle of miRDeep2 (16) as the most frequently used prediction tool for miRNAs, we implemented an own predictor with an extended feature set including our previously developed prediction score (11). Furthermore, we implemented functionality to report the presence of miRNA motifs to the user (25–27). MiRMaster allows to search for novel miRNA candidates, to quantify miRNA expression, to identify isoforms and variants of miRNAs. Another feature of miRMaster is the mapping of non-human small RNA reads against the NCBI RefSeq collection of bacterial and viral genomes (28), thereby allowing the detection of contaminations, infections or exogenous miRNAs. To allow the analysis of targets regulated by miRNAs, we implemented Application Programming Interfaces (APIs) to available web-based tools for considering the targetome (miRTargetLink (29)) and to carry out miRNA set enrichment (miEAA (20)).

Since different research groups measured various specimens using different experimental protocols and bioinformatics pipelines and not all data stored in a central repository, a redundancy between the studies exist. Besides the miRNAs in the miRBase, and specific studies mentioned before, several comprehensive analyses (e.g. Londin *et al.* (30), Backes *et al.* (11), Friedländer *et al.* (31), Jha *et al.* (32)) propose hundreds to thousands of new miRNAs. To detect as many as possible miRNA candidates we performed a comprehensive analysis of 1836 data sets containing 20 billion reads.

MATERIALS AND METHODS

Sample collection

As case study we analyzed an in-house NGS miRNA sample collection of 1097 samples from blood and blood cell components (33–39). Further we downloaded 739 samples from four series of the GEO database (40): GSE64142, GSE53080, GSE49279 and GSE45159. All samples have been sequenced using Illumina Next-Generation sequencing. Table 2 presents an overview of these samples including a description, number of samples, number of reads and file size.

Positive miRNA dataset for training miRMaster

A straightforward positive dataset would consist of the complete miRBase (1). However, others and we have observed that miRBase may contain false positives, especially in the last versions (41). Therefore, we selected all miRNA precursors from miRBase 1 to 7 and all precursors of miRNAs containing strong experimental evidence in the miRTarBase (42), leading to 487 high-confidence positive miRNAs. We defined precursors by their 5′ and 3′ mature miRNAs, i.e. they start with the first base of the 5′ miRNA and end with the last base of the 3′ miRNA. For miRBase precursors that had only one form annotated we derived the other from its hairpin, as described for our prediction algorithm. Therefore, our predictions are independent of the size of the stem loops provided in miRBase.

Negative miRNA dataset for training miRMaster

Choosing an appropriate negative dataset is a challenging task, since miRNAs can be located anywhere in the genome (43). A correct negative dataset plays an important role for the creation of a well-trained classifier. Overall, since only a small fraction of the genome and of sequences that form hairpins are actually precursors, we built five different sets to cover as many potential wrong predictions as possible. The different negative datasets were derived from separate assumptions and combined for our training procedure. The first dataset was built to cover predictions, where one actual miRNA is contained in the predicted precursor but the other miRNA is wrongly annotated. We assume that real precursors do not overlap. It was created by splitting in half all known stem–loops from miRBase that contained two annotated mature miRNAs. We adjusted the length to the original stem-loop by including the flanking regions. To determine the positions of the miRNAs in the two new pseudo precursors, we kept the original miRNAs and derived the other based on it, as in our prediction algorithm. This dataset was composed of 298 precursors. The second dataset was created to cover predictions that could stem from protein coding sequences of genes without known alternative splicing events. It was derived from the widely used pseudo precursor set built by Xue *et al.* (44). We first kept only sequences that aligned perfectly to the latest assembly of the human genome (hg38). Then we segmented these sequences to enable the computation of segment specific features. Therefore, we determined the position of one of the pseudo miRNAs by assigning it to the segment with most base pairs, having a length of 20 nucleotides and non-overlapping with the loop region. The other was derived from it, as in our predicting algorithm. The resulting set contained 3916 pseudo precursors. The third dataset was created to cover predictions that could arise from stem-loops of other ncRNAs. It was shown by others (45) that for a very small portion of all known miRNAs this could actually be the case. However, due to their low number and the false positives largely outweighing the true positives we considered this set to be useful to reduce the false positive prediction rate. The dataset was derived from Rfam (46) (release 11) and composed of 3342 negative precursors. We considered all human ncRNAs that were not miRNAs and derived pseudo precursors by retaining only those that could be partitioned into 5′, 3′ and loop parts. The fourth dataset was created to account specifically for predictions that would pass the filtering steps in our algorithm, but which would overlap with other ncRNAs. It is in fact an extension of the third dataset. We derived 4031 pseudo precursors by running our prediction on 705 in-house samples and keeping only those that passed all filtering steps but overlapped with other ncRNAs of Rfam. The fifth dataset was created to account for predictions that were not covered by the other negative datasets. It was derived from early predictions performed by our algorithm (trained on the other four datasets) on our in-house samples. This set addresses specifically predictions where the miRNAs contained many repeated bases and further, miRNA duplexes with high normalized free energy and precursors with high normalized free energy. We kept all predictions that displayed evidence for being false positives, i.e. precursors with miRNAs containing at least seven consecutive A or U or 8 C or G. Further we kept all with a normalized ensemble free energy of over –0.15 kcal/mol*nt or with a normalized duplex minimum free energy of over –0.15 kcal/mol*nt. The cutoffs were determined empirically by analyzing the distribution of the properties of known precursors. This led to 797 additional negative miRNAs. For the first four datasets we further retained only those pseudo precursors without bifurcations, with at least 50% paired bases between the 5′ and 3′ pseudo miRNAs and with a 5′-3′ miRNA length difference of at most 10. The combination of all negative datasets resulted in 12 384 pseudo precursors, which are listed in Supplementary Table S2.

Independent test sets for evaluating miRMaster

To validate the performance of our model we created two additional independent test sets. The first set was composed of human precursors of MirGeneDB (10) that were not used in our training process, resulting in 129 precursors. For the pseudo precursors we selected all sequences that were annotated as human precursors in earlier miRBase versions (1–20) and that were not duplicates or merged with known precursors. This resulted in 28 sequences, of which 6 were discarded by our algorithm when trying to determine a valid corresponding second miRNA arm. In addition, we created a second set composed of mouse precursors of MirGeneDB that had different sequences than our training precursors, resulting in 350 precursors. We selected the negative set analogously to the first negative set from early annotated mouse precursors, leading to 65 sequences. We mapped those sequences against the mouse genome (mm10) and removed all sequences which were not found or found at multiple positions. Of the remaining 56 sequences, 11 were discarded by our algorithm when trying to determine a valid second miRNA.

Features of miRMaster for predicting novel miRNAs

We created a feature set composed of 216 properties, based on 186 existing features described in (44,47–51) and 30 novel features. Novel features included our previously developed novoMiRank score (11), open/close parentheses and unpaired nucleotides in all thirds of a precursor, 5′-3′ miRNA duplex minimum free energy, the number of base pairs in the 5′ and 3′ miRNAs and in-between, and the nucleotide ratio of the 5′ and 3′ miRNAs. Supplementary Table S1 lists all features including a brief description, their runtime impact and the *P*-value resulting from a two sided Wilcoxon rank-sum test after Benjamini–Hochberg adjustment for multiple testing (52) (alpha = 0.05) on our positive and negative datasets.

Classifier selection for predicting miRNAs

To obtain the best classifier for our positive and negative dataset in terms of specificity and sensitivity we evaluated 180 different combinations of feature scaling, subset selection and classification methods using the scikit-learn Python toolkit (53), as shown in Supplementary Table S9. Since a large fraction of features can be computed in minimal time while very few features take very much computing time we built two models: one is based on all features and one based on the features with low runtime. For each combination we tuned the classifier's hyper-parameter via particle swarm optimization towards maximum ROC AUC, resulting in a total of 130,105 models. From those we then selected all models that performed at least as good as the best 25% according to ROC AUC, Precision-Recall AUC, sensitivity, specificity and Matthews correlation coefficient (MCC). The final model was chosen according to the highest $F_{0.5}$ measure. Supplementary Figure S15 sketches this process.

Input data of users to miRMaster

Since our ambition was to facilitate comprehensive miRNA analysis for all researchers, we implemented upload functionality for FASTQ files that are processed and compressed in the browser before being sent to the server. Thus, no additional software installation that compresses the files on the user's computer is needed. This feature is supported by only few tools, such as MAGI (54). Further we provide support for gzip compressed FASTQ files, since they are the typical storage format of sequencing files, thereby obviating the need to decompress files before inputting them to miRMaster.

Preprocessing

Before sending the input files to our server we perform three preprocessing steps consisting of adapter trimming, quality filtering and read collapsing. Adapter trimming is performed via fuzzy string matching and can be customized by the user. We allow one mismatch and require an overlap of at least 10 nucleotides with the read per default. Further the user has the possibility to trim leading and trailing *N*, discard reads containing any remaining *N* and remove reads shorter than a specific size. For the quality filtering step, we re-implemented the sliding window filtering approach used by Trimmomatic (55). This allows reducing the amount of data sent by up to 99.9% (depending on the sample specimens). To take advantage of multi-core processor capabilities we use JavaScript web workers to allow the preprocessing of multiple files at the same time.

Mapping to various ncRNA databases

We map the collapsed reads using Bowtie (56) and allow per default no mismatches against human rRNAs, snRNAs, snoRNAs, scaRNAs and lincRNAs of the Ensembl noncoding RNA database (release 85) (57), against piRNAs of piRBase (1.0) (58) and tRNAs of GtRNAdb (59). This allows the user to easily verify if the distribution of reads is as expected or to investigate specific RNAs. To allow the user to investigate specific ncRNAs we provide detailed expression counts for all ncRNAs we are mapping against, as well. The expression is determined by the number of reads mapping to a specific sequence using Bowtie. Further we report the mapping of reads against the human miRBase (version 21), which can be used to estimate the potential of finding novel miRNAs in the samples.

Mapping to reference

Mapping the collapsed reads to the reference genome is performed using Bowtie. Analogous to miRDeep2 (16), we require no mismatches in the first 18 nucleotides and discard reads that map to over five different locations.

Precursor excision, segment determination and filtering

The precursor excision, segment determination and filtering according to their structure and signature is performed analogous to miRDeep2. Briefly, local maximum read stacks in downstream windows of 70 nucleotides are searched and two precursors excised from each stack. The secondary structure is computed for each precursor using RNAfold (60). The maximum read stack represents one miRNA of the precursor. The other miRNA is determined by the paired sequence on the other arm with a 2-nucleotide overhang. Filtering steps are composed of a structure and signature filter. The secondary structure is required to have no bifurcations, a minimum percentage of base pairs in the highest expressed miRNA of 60% and a length difference of both miRNAs of at most five nucleotides. The signature is checked by mapping all reads with at most one mismatch against all excised precursors. At least 90% of all reads need to map to either a miRNA or in between, thereby discarding reads that do not map according to Dicer processing. All these thresholds can be customized in the web interface.

Feature computation and prediction

After the potential precursors have been excised and filtered we compute their feature values and perform the prediction using our classifier as described in previous parts of the Materials and Methods section.

Prediction merging and global signature filtering

Once the predictions for all samples have been performed we merge the resulting potential precursors in order to avoid multiple predictions shifted by only a few bases. Therefore, we group all precursors that differ by at most 10 positions and keep the one that was found in most samples. To make use of additional information provided by multiple samples we first normalize the expression of each read of each sample to reads per million (RPM) and sum up identical reads. Then we map the normalized reads of all samples against the merged predictions and score their signature. We weight each read using the following formula

$$
\begin{aligned}
&score\,(read)\\
&= total\_RPM(read) \cdot length\,(read) \cdot \sqrt{\frac{occuring\_samples\,(read)}{\#total\_samples}}
\end{aligned}
$$

Thereby, we penalize reads that occur in only few samples while giving more weight to longer reads. Reads mapping with mismatches are penalized per default by a dividing factor if they occur in at most 10% of all samples (but at most 10 samples). The dividing factor is the limit of occurring samples minus 1, but at least 2. We then remove all predictions that have a signature with an inconsistent dicer processing read portion representing at most 20% of the total score.

Categories of new miRNAs

We assign to each predicted precursor one of six categories. (1) *Known*: when the prediction is overlapping with a miRBase entry and both miRNAs are overlapping with known miRNAs by at least 75%. (2) *Shifted known*: when the prediction is only partially overlapping with miRBase and only one miRNA is overlapping by at least 75% with a known miRNA. (3) *One annotated*: when the prediction is overlapping with a miRBase entry, but only one miRNA is annotated for that entry and this one is overlapping by at least 75%. (4) *Dissimilar overlapping*: when the prediction is overlapping with a miRBase entry, but the miRNAs are not overlapping with the annotated ones. (5) *Half novel*: when the prediction is not overlapping with any miRBase entry, but contains at least 75% of one known miRNA. (6) *Novel*: when the prediction is not overlapping with any miRBase entry and does not contain any known miRNA.

Prediction flagging of other ncRNAs

In order to reduce the number of potential false positives, we map the predicted precursors to the Ensembl human non-coding RNA database (release 85) and to NONCODE 2016 (61) using BLAST+ (62) and flag them accordingly when matches are found. Further we map against the whole miRBase (v21) to highlight similar miRNAs in other species. Mappings are valid when over 90% of the aligned sequences overlap and at most one mismatch is present.

Quantification of known and novel miRNAs, isomiRs and mutations

The quantification of known and novel miRNAs is performed analogously to miRDeep2. Reads are mapped against the precursors using Bowtie while allowing one mismatch. The counts are reported for all reads overlapping the annotated miRNAs in a window of up to two nucleotides upstream and five nucleotides downstream. IsomiRs are detected by mapping against the precursors using Bowtie while tolerating two mismatches. We allow up to two non-template additions to the 5′ and 3′ ends and up to one mismatch in between. We also allow a variability of two nucleotides at the 5′ end and of five nucleotides at the 3′ end per default. When detecting mutations, we focus on single nucleotide substitutions. The mapping and counting is performed the same way as the quantification, however miRNAs with mutations are explicitly counted.

Exogenous read mapping

We map non-human reads (all reads that did not align to the human genome with at most one mismatch) to all 7556 bacteria and 7026 virus sequences of NCBI RefSeq (28) release 74 and report the number of perfectly mapping reads.

Reads mapping to bacteria or viruses can indicate exogenous miRNAs, but also reagent contamination or diseases such as sepsis.

Motif detection

Recently five miRNA motifs have been reported, namely the UG, UGU/GUG, CNNC (25), GHG (26) and GGAC (27) motif. We report for each prediction the present motifs, allowing matching up to two nucleotides upstream or downstream of the expected motif position.

Usability

To analyze NGS miRNA samples with miRMaster, the user needs to provide sequencing files in FASTQ format (uncompressed or gzip compressed) without barcode sequence and the 3′ adapter used in the library preparation. After clicking on the 'Launch experiment' button on the homepage or in the navigation bar, the user will be guided through three steps. During the first one, one should name the experiment and also optionally provide an e-mail address to receive a notification as soon as the analysis of the uploaded samples is done. During the second step the user needs to specify the used 3′ adapter and has the opportunity to fine-tune the parameters of the analysis. The third step consists of the upload of the sequencing files. If the samples stem from multiple cohorts, groups can be specified by either clicking on the 'Add second group' button or by uploading a tab separated sample-to-group file. Once the files are chosen and the user has clicked the 'Launch' button, the data will be preprocessed and sent to the server. The preprocessing progress is shown directly on the web page whereas the server progress can be followed in real time by clicking the 'Follow' button. This will open the experiment status page in a new tab, where the user will be able to track the progress of the analysis of all uploaded samples. Real-time web reports are provided for each sample that has been uploaded, allowing to directly inspect the data. These reports provide information on the preprocessing, mapping, quantification and prediction steps. As soon as all samples have been analyzed, the results can be downloaded and an overall web-report is created with a link to it on the top of the status page.

Validation using custom microarray

To perform a first pass iteration and to minimize the risk of false positives due to either NGS artifacts or low sample quality containing many degraded RNAs we designed a custom microarray containing all human miRNAs from the miRBase, the miRNAs from the study by Londin *et al.* (30) as well as over 5000 miRNAs from the present study. Among our predicted miRNAs we selected only those expressed in at least 50 samples which were not flagged as similar to other ncRNAs. The final microarray contained 11 866 miRNA candidates that have been measured each in 20 replicates (237 320 features per sample).

In order to measure the expression of the novel miRNAs in different human cells and tissues, we compiled a set of eight different human RNA samples: we purchased human total RNA samples from lung, brain, kidney, testis and heart tissues from Life Technologies (Cat. No. AM7968, AM7962, AM7976, AM7972 and AM7966, respectively) and the human miRNA reference kit from Agilent Technologies (Cat. No. 750700), that represents a pool of several human tissues and cell lines. Furthermore, we used a PAX blood RNA pool and a plasma RNA pool. The PAX blood RNA pool comprised of 11 blood samples collected in PAX gene tubes and purified with PAXgene Blood miRNA Kit from Qiagen according to manufacturer's instructions. Blood samples derived from four lung cancer patients, two Alzheimer's Disease patients, two patients with Wilms Tumor, and three healthy donors. The plasma RNA pool comprised of 10 plasma samples from healthy donors and was isolated using miRNeasy Serum/Plasma Kit after manufacturers recommendation with minor adaptations. To ensure sufficient RNA precipitation, we added 1 μl 20 mg/ml glycogen (Invitrogen) in the precipitation step. RNA concentration was measured using Nanodrop (ThermoFisher). RNA quality was assessed using Agilent Bioanalyzer Nano kit (for all tissue derived RNAs) or Small RNA kit (for the plasma sample).

The expression of 11 866 miRNAs and miRNA candidates was determined using the customized Agilent human miRNA microarrays. As input we used 100 ng total RNA as measured in Nanodrop for all tissue derived RNAs, and 1 ng miRNA as measured using Bioanalyzer Small RNA chip for the plasma sample. Using Agilent miRNA Complete Labeling and Hyb Kit after manufacturer's instructions, RNAs were dephosphorylated and labeled with Cy3-pCp. Labeled RNAs were hybridized to the custom microarrays for exactly 20 hours at 55°C. After hybridization, arrays were washed for 5 min in each Gene Expression Wash Buffer 1 (room temperature) and 2 (37°C). Subsequently, arrays were dried and scanned in an Agilent microarray scanner (G2505C). Expression data was extracted using Agilent feature extraction software. Downstream processing of signals has been carried out with R (version 3.2.4). Specifically, for clustering the expression intensities hierarchical clustering using the Euclidean distance has been performed as implemented in the Heatplus package.

To enable other researchers to repeat the experiments and to perform measurements on own samples, the microarrays that can be used analogously to standard Agilent microarrays using the Agilent protocols and SureScan platform, will be distributed by Hummingbird Diagnostics (Heidelberg, Germany) in three versions: human-mirna-candidate(full) containing all miRNA candidates from this study; mirna-candidate(detected) containing all miRNAs positive in any experiment of this study; mirna-candidate(blood) containing all miRNAs that have been detected in blood or serum.

RESULTS AND DISCUSSION

The aim in developing miRMaster (www.ccb.uni-saarland.de/mirmaster) was to implement a comprehensive tool for the analysis of miRNA NGS data sets. Starting from raw or compressed FASTQ files with billions of reads and gigabytes of data, miRMaster allows a wide variety of miRNA analyses. The complete workflow is described in detail in the Methods section and sketched in Figure 1. A brief de-

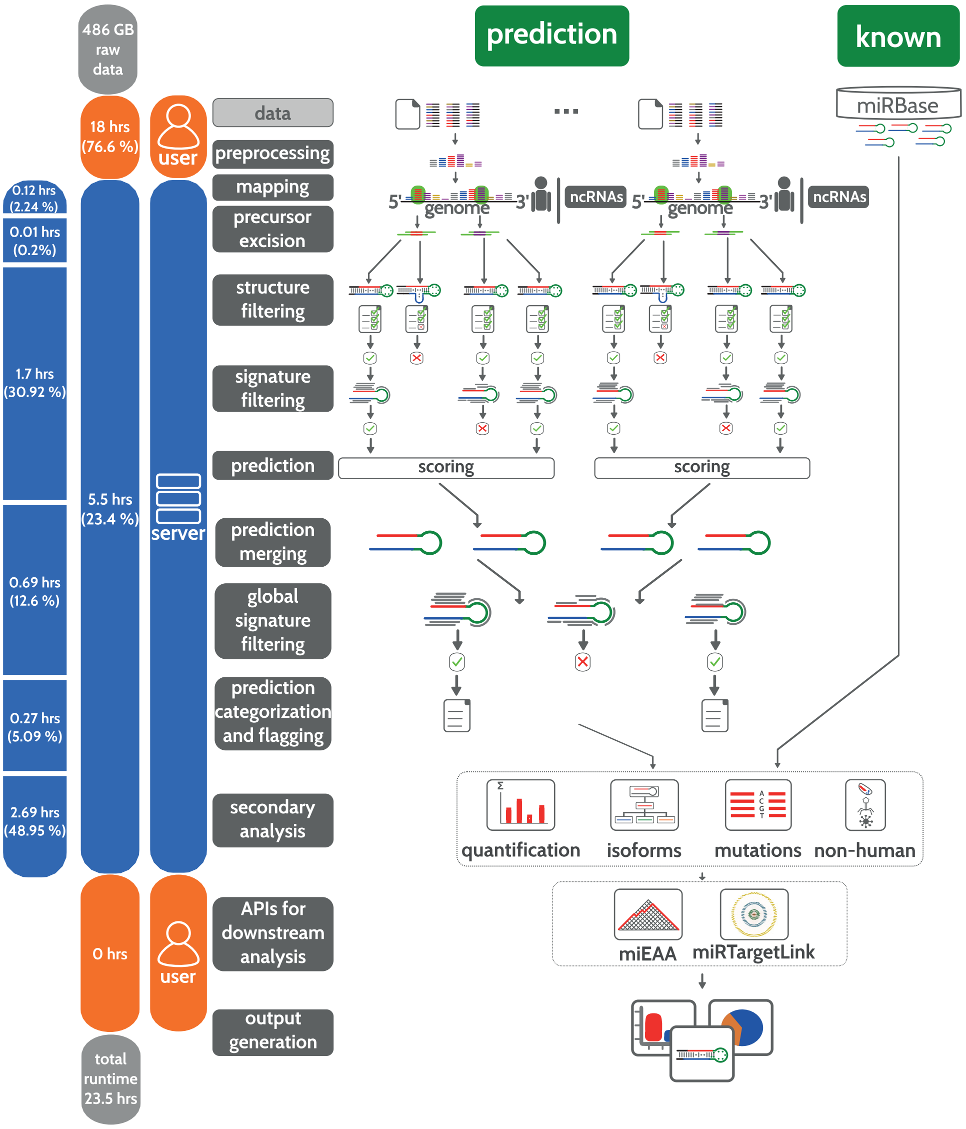


Figure 1. Schematic workflow of miRMaster. The bar at the left shows the runtime impact of each step. Steps performed by the user are shown in orange and steps performed by the server in blue.

scription on the usability of miRMaster is available in the Methods section.

In the following, we first focus on the performance of the novel algorithm for the prediction of new miRNAs. In total, we investigated 1097 miRNA NGS data sets containing 15 billion reads within a 486 GB file size and compare the miRMaster results – in terms of performance and runtime—to those of miRDeep2 using the same data sets. We next provide a detailed description of the different components of our miRNA NGS analysis framework and their application to the above-mentioned data set. Then we report a coarse description of the human miRNome by predicting small RNAs from 1836 data sets with 20 billion reads. Finally, we analyze the expression of potential miRNA candidates using custom microarrays.

Evaluation of miRNA features

In contrast to most other comparable tools, our miRNA prediction relies on a broad set of features that are derived both from precursor sequences and from their mature forms. These features are considered as weak learners as each feature has a limited impact on the overall decision to classify or declassify a new miRNA as true miRNA. The feature set consists of 216 single features including nucleotide composition, secondary structure and others (the full list is available in Supplementary Table S1). To gain first insight into the discrimination power of single features we derived a positive miRNA precursor set from early miRBase (63) versions and from targets with strong experimental evidence in miRTarBase (42) (487 precursors), as well as a negative miRNA precursor set from various sources (12 384 negative precursors). A detailed explanation on the creation of these sets can be found in the Methods section (the sequences and locations of both sets are shown in Supplementary Table S2). We calculated the significance of all features by comparing both sets via Wilcoxon rank-sum tests. The performance of the 216 features is listed in Supplementary Table S1. The smallest significance value (10^{-219}) was calculated for the minimum free energy index 1. Following adjustment for multiple testing, 158 of the 216 features remained significant ($P < 0.05$). Since our analysis pipeline is designed to support the evaluation of large data collections of up to several thousand samples, performance in runtime of feature calculation is of importance. We grouped all features in three different runtime categories with the fastest category containing features with 10,000-fold decreased runtime as compared to the slowest features. Supplementary Figure S1 shows the negative decadic logarithm of the *P*-values for features in the three categories. Since the two fast categories already contained 54 and 86 significant features, respectively, we evaluated their combined information content for predicting miRNAs. We derived classifiers not only from the complete feature set, but also from the fast features set only. Prior to classifying miRNAs based on the features we evaluated the redundancy of the features selected. As shown in the correlation heat map in Supplementary Figure S2 many of the features were redundant.

Classification of precursors

For combining the predictive power of the weak learners we applied different feature selection and classification approaches. We selected a large variety of classifier and feature selection approaches, since there is no ‘one size fits all’ approach and our goal was to build a model that performs best on our datasets. Each of the tested classifiers and feature selection approaches have their strengths and weaknesses (e.g. SVMs with different kernels are suitable for different kinds of separation spaces). Since several single features show low discriminatory power (Supplementary Figure S1) and many features are correlated to each other (Supplementary Figure S2) it is important to define feature subsets that allow to classify or declassify a new miRNA precursor as true precursor. Different scaling and feature selection methods can have substantial effects on the used classifier. Therefore, we performed an exhaustive analysis of all combinations. We evaluated 130 105 different combinations of feature selection and classifiers using repeated stratified 5-fold cross validation. Even with the cross-validation, the evaluation of so many different classification attempts may lead to overoptimistic results. To address this problem, we performed permutation tests. The evaluation of the key performance criteria in Table 1 shows that almost all classifications were highly accurate. The area under the receiver operating characteristic curve (ROC AUC) highlights median performance of 99%, with the 90% quantile of all approaches being at 99.5% and more impressively the 10% quantile being at 95.8%. In consequence, 90% of all 130 105 tested classifiers had an AUC exceeding 95.8%.

For both, the complete and the fast feature set AdaBoost outperformed the other models with an AUC of 99.6%, a specificity of 99.9% and a sensitivity of 86.9% for the complete feature set, and an AUC of 99.4%, a specificity of 99.9% and a sensitivity of 83.4% for the fast feature set. The selected AdaBoost classifier by itself selects only features known to improve the prediction and is therefore well suited for our broad set of features. This comparison demonstrates that the performance of the fast feature set is only marginally weaker than the performance of the full feature set. Nonetheless, we evaluated the performance of these two models and carried out stratified 5-fold cross-validation with 1000 repetitions each. The same approach was done with 1000 permutation tests each. As shown in Supplementary Figure S3, random test performance did not compare to the true performance in any of the cases and cross validation performance was stable and good in all cases. This further suggests that the composition of the cross-validation splits plays no major role for the model performance. In addition to the cross-validation performance we evaluated our model with the fast feature set on two independent test sets. A description of the independent test sets can be found in the Materials and Methods section. The first test set was composed of 129 human precursors and 28 human pseudo precursors. On this set our model reached a sensitivity of 82.9% and a specificity of 100%. The second test set contained 350 mouse precursors and 56 mouse pseudo precursors and resulted in a sensitivity of 81.4% and a specificity of 98.2%.

Evaluation of prediction from 1097 miRNA NGS samples

Having evaluated the performance of our classifier on the positive and negative training set we applied the models to 1097 in-house data sets (33–39). These contain 15 billion reads in a total file size of 486GB (see Table 2). Again, we first compared the fast feature set versus the complete feature set. The prediction was carried out for each sample individually. They were then merged and filtered according to their global read signature. The differences between the models regarding known miRNAs were minimal with both models discovering 900 precursors, while 55 additional were uniquely found in the fast model opposed to 34 in the full model, as shown in Supplementary Figure S4. As for the novel miRNAs both models discovered 10 651 precursors. We then compared the unique predictions of both models in regard to their mean probability, novoMiRank score and the number of samples they were predicted in. We found that their mean scores and the mean number of samples they were predicted in were very similar (score of 1.18 for the complete model, 1.19 for the fast one; predicted in 7.5 samples for the complete and 7.6 for the fast model). However, we noticed also that for both sets the majority of the differing predictions were near the decision boundary with a mean probability below 60% (in contrast to an average of 70% for the common set), meaning that these predictions were among the less likely precursor miRNA candidates. Therefore, since both models performed very similarly, ex-

Table 1. Cross validation performance

	Specificity	Sensitivity	Accuracy	NPV	Precision	ROC AUC	$F_{0.5}$
Median	99.78%	70.62%	98.61%	98.90%	91.37%	98.98%	85.10%
90% quantile	99.91%	82.35%	99.18%	99.34%	95.61%	99.50%	91.81%
10% quantile	99.44%	45.17%	97.41%	97.97%	73.60%	95.85%	64.80%
AdaBoost (all features)	99.98%	86.85%	99.51%	99.51%	99.54%	99.58%	96.71%
AdaBoost (fast features)	99.98%	83.37%	99.38%	99.38%	99.26%	99.39%	95.60%

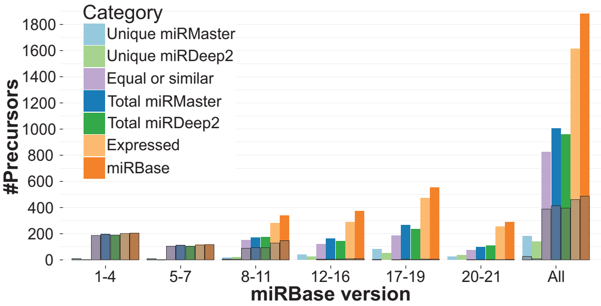


Figure 2. Distribution of recovered known miRBase precursors using miRMaster and miRDeep2. Predicted precursors are regarded as similar if they overlap by at least 90%. The black boxes show the number of precursors contained in the training set of miRMaster.

cept for the less likely candidates, we further focused on the fast model, due to its runtime advantage.

Comparison between miRMaster and miRDeep2

To further evaluate the performance of miRMaster we compared its predictions with the predictions of miRDeep2, one of the central programs for miRNA discovery. In detail, we ran miRDeep2 with default parameters on our 1097 NGS samples and merged the overlapping precursors predicted by miRDeep2 by retaining the precursors predicted in most samples. The same procedure was applied for miRMaster. A more detailed description of the different analysis steps can be found in the Methods section.

As shown in Figure 2, miRDeep2 recovered 59.5% of the known miRBase (version 21) precursors detected by quantification while miRMaster found 62.3% of them. Further, miRMaster consistently recovered more precursors from our training set than miRDeep2 (in total 414 versus 396). Specifically, 181 precursors were exclusively found by miRMaster and 138 by miRDeep2 as shown in Supplementary Table S3. Figure 2 shows that both tools perform especially well in earlier miRBase versions with both tools reporting nearly all precursors up to miRBase version 7. Precursor miRNAs exclusively detected by miRDeep2 are mainly found in later miRBase versions and contained only 7 precursors of miRNAs with strong experimental evidence for targets in miRTarBase. By contrast miRMaster detected 21 precursors in later miRBase versions with strong experimental evidence for targets in miRTarBase. These results might be biased since our models contain many more features and are trained using human high-confidence miRNAs on the one hand, and many miRNAs in later miRBase versions have already been reported by miRDeep2 on the other. Overall, the data suggest that our classifier identifies more known miRNAs and especially more of the strongly confident miRNAs.

To present a realistic comparison in runtime of miRMaster and miRDeep2, we measured execution time on the same infrastructure starting from pre-processed data. The computations were performed on a node with four AMD Opteron 6378 (4 × 16 cores totaling 64 cores) at 2.4 GHz and 512GB DDR3-RAM. MiRDeep2 required 102.5 h (4.4 days) without PDF generation (usually increases the runtime by 40% and produces reports for each known and predicted precursor). The respective steps of miRMaster required only 5.5 h which is a 19-fold decrease in runtime compared to miRDeep2. The difference is especially notable since miRMaster performed many additional analyses such as prediction of isoforms, variants in miRNAs and others. This difference in runtime is explained by the computed features and by different implementations. While miRDeep2 is implemented in Perl, miRMaster relies on a more efficient implementation in C++ for substantial parts of the program. One example is the precursor excision step, a reimplementation of the miRDeep2 Perl code in C++. This part of the program is roughly 40-fold faster in miRMaster as compared to miRDeep2.

A detailed break-down of the runtime in the different steps is presented in Supplementary Figure S5. The reads are mapped against miRBase and multiple other ncRNA databases (1.52% of the runtime) and to the human genome using Bowtie (56) (0.72% of the runtime). The afore mentioned precursor excision step requires 0.2% of the runtime. The following steps that are central for miRMaster include precursor segmentation, filtering, feature computation and prediction, altogether requiring 30.92% of the runtime. The predicted miRNA precursors from different samples are subsequently merged and filtered according to the read profiles of all samples (12.60% of the runtime). The following assignment to one of six categories ‘known’, ‘shifted known’, ‘one annotated’, ‘dissimilar overlapping’, ‘half novel’ or ‘novel’ requires 0.75% of the runtime. For the prediction flagging step, ncRNAs from Ensembl (57), lncRNAs from NONCODE (61) and known miRNAs from miRBase are mapped against the precursors (4.34% of the runtime). Finally, different secondary analyses are carried out on known and novel miRNAs, including quantification, which is again a reimplementation of miRDeep2, detection of isoforms and single base mutations. These steps, including the mapping of non-human reads to a collection of 7556 bacteria and 7026 viruses of NCBI RefSeq, permitting the detection of potential exogenous miRNAs, require in total 48.96% of the server runtime.

Table 2. Composition of all 1836 NGS samples

Source / Description	#Samples	#Reads	Compressed File Size
CNS lymphoma patients and controls (in-house)	44	884 Mn	25GB
Alzheimer patients and controls (in-house)	203	3.4 Bn	114GB
Cardiovascular disease patients and controls (in-house)	485	6.9 Bn	205GB
Multiple sclerosis patients and controls (in-house)	217	1.2 Bn	44GB
Blood cell fractions from healthy donors (in-house)	148	3.3 Mn	98GB
GSE64142 (monocyte-derived dendritic cells upon bacterial infection)	116	1.4 Bn	43GB
GSE53080 (myocardium, plasma and serum in heart failure patients)	185	925 Mn	36GB
GSE49279 (adrenocortical tumors)	78	1.2 Bn	34GB
GSE45159 (adipose tissue)	360	786 Mn	24GB
Sum	**1836**	**20 Bn**	**623GB**

Web-based analysis using miRMaster

With the development of miRMaster we aimed to provide a comprehensive web-based toolbox for an all-in-one miRNA analysis. In detail, the web-based tool has to (a) enable the analysis of HT-sequencing raw data without installing any software, even for data sets in the range of dozens of gigabytes; (b) perform the most common and further specialized analyses in an integrative manner; (c) return the results in a manner to be used for identifying interesting hits and for publication purposes by wet-lab scientists. These analyses are carried out in a fully integrated manner. From the raw data input (1097 compressed FASTQ files, 486GB) to final results for all calculations, miRMaster required 23.5 h. Data upload at client side was performed on an Intel Core i5–5200U Notebook with 12GB DDR3-RAM using Mozilla Firefox 48 and required most of the time (18 of the 23.5 h), while the analysis of pre-processed data took only 5.5 h. At client side, FASTQ files are first pre-processed (adapter trimming, quality filtering, read collapsing) and subsequently uploaded. The functionality is implemented in JavaScript such that no software has to be installed by the user. The runtime of this step may vary based on the equipment at user site and the bandwidth for data upload. Real world tests have demonstrated that studies including e.g. 50–100 samples can be evaluated in well below 5 h.

Evaluation of variations in miRNAs by miRMaster

First, we investigated the mutation frequency. For each known miRNA of each of the 1097 samples we searched the number of single base mutations. To reduce a bias depending on the coverage we considered only miRNAs and their variants covered by at least 30 reads in 100 samples. Out of 2147 detected miRNAs 333 fulfilled the criteria. Supplementary Table S4 lists the mutations found in all miRNAs. Overall the largest number of variants was discovered for hsa-miR-486-5p, which is abundantly expressed across all samples with two precursors. However, for the majority of miRNAs the number of variants is low with most miRNAs having two or less variants (67.3%). For some miRNAs, such as hsa-miR-6131 the unmutated form was almost never detected and only variants with mutations at position 8 and 14 were found. Another example is hsa-miR-1260b with the most abundant form showing an A→G mutation at position 8 (Supplementary Figure S6). However, for most miRNAs (91.6%) the wildtype was most expressed. Our results suggest that only a small set of miRNAs is frequently affected by mutations e.g. due to RNA editing. The low number of mutations is to be expected, since mutations, especially in the seed region, are likely to highly affect the miRNA regulation network.

Next, we calculated for each known miRNA the number of isoforms, analogously to the steps performed for the detection of single base mutations. After applying the abovementioned filter criteria, we found 277 miRNAs isoforms that are listed in Supplementary Table S5. As for the mutated miRNAs we found the by far largest number of isoforms for hsa-miR-486-5p, which is highly expressed in blood. In consistence with the single base mutation results, the number of variants is low for the majority of miRNAs with most miRNAs (53.8%) showing four or less variants. For most miRNAs (71.5%) we detected the canonical form as annotated in miRBase. The miRNA with most variants and without canonical form was hsa-miR-107. As shown in Supplementary Figure S7, the most expressed form of hsa-miR-107 with a median of over 60% was trimmed by four nucleotides from the 3′ end, resulting in a miRNA with 19 nucleotides. Further, we frequently observed a lack of a dominating isoform over all samples, as for example for hsa-miR-29a-3p (Figure 3). This is consistent with the idea that isoform expression varies depending on the context, such as the cell type, time or population. Since the canonical form was most expressed in only 33.6% cases, isomiRs apparently play an essential role in miRNA function.

Comprehensive version of the human miRNome

Currently, the total number of human miRNAs is controversially discussed. While miRBase currently contains 2588 human mature miRNAs (version 21), several studies propose even larger sets (e.g. Londin *et al.* (30), Backes *et al.* (11), Friedländer *et al.* (31), Jha *et al.* (32)). There exist two major challenges. First, the different miRNA sets are partially overlapping or contain miRNAs shifted only by few bases, adding a substantial redundancy. Second, the miRBase contains many false positive miRNAs, especially in later versions.

Using miRMaster we attempted to generate a coarse description of the human miRNome, i.e. we wanted to describe as many putative miRNA candidates as possible, being well aware that false positives are included (e.g. tRNA fragments, piRNAs or artifacts). This collection of potential candidates can be used to minimize further redundancy in upcoming high throughput studies.

Thus, in addition to our in-house NGS samples, we collected 739 samples from GEO (40), resulting in 1836 NGS

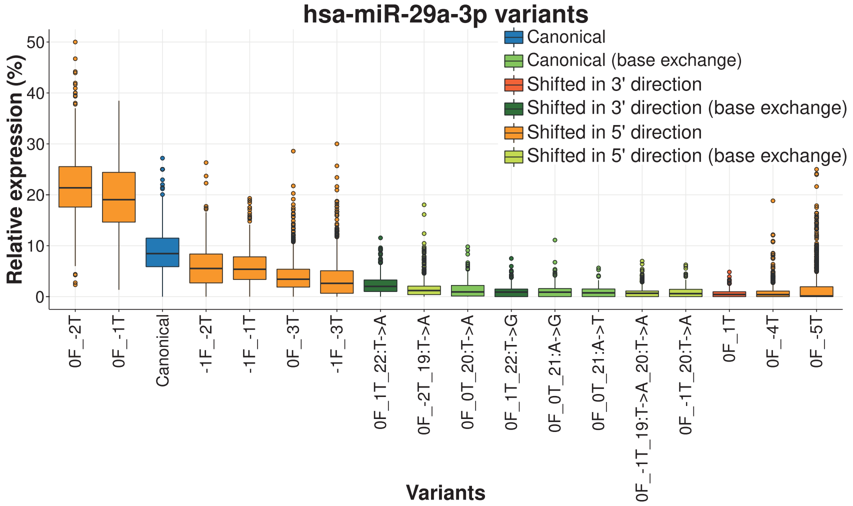


Figure 3. Isoform distribution of hsa-miR-29a-3p. Only variants appearing with an evidence of at least 30 reads in 100 samples are shown on the x-axis. Only reads occurring at least 30 times in a sample are shown for the relative expression to avoid large outlier due to low raw expression. Isoform notation: the number before F stands for the distance to the canonical 5′ end, in 5′-3′ direction (i.e. positive for trimmed, negative for extended); the number before the T stands for the distance to the canonical 3′ end (i.e. negative for trimmed, positive for extended). The canonical form is the third most frequent one and is highlighted in blue. Variants without base exchange are frequently shorter or shifted in the 5′ direction (orange), those with base exchanges match either the star/stop of the canonical miRNA (green) or are shifted slightly to the 5′ (light green) or 3′ (dark green) direction.

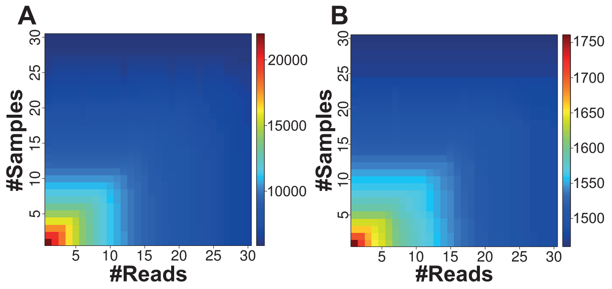


Figure 4. Distribution of the number of expressed precursors according to an evidence in a minimum number of samples and a total minimum number of reads. (**A**) The distribution of the number of expressed novel precursors. (**B**) The distribution of the number of known precursors.

samples (Table 2), and predicted novel miRNAs on those samples. The run resulted in 21 996 novel predicted miRNA precursors that are listed Supplementary Table S6. Those predictions can be inspected on the miRMaster webpage and downloaded as FASTA format. As shown in Figure 4A, most of the novel precursors were weakly expressed and in few samples. Considering only miRNAs with an expression in at least 30 samples reduced the number of predictions to 5845. As displayed in Figure 4B, the known precursors of miRBase (version 21) seem to be less affected by the augmenting number of samples or reads. Supplementary Figure S8 shows the number of expressed known and novel precursors according to their expression in multiple samples. The number of novel precursors decreases exponentially and faster than the known precursors with increasing number of required samples. This suggests that the majority of the commonly expressed miRNome is already known and that mainly tissue specific, time specific or other context specific miRNAs remain to be discovered.

Precursors of known and new miRNAs are evenly distributed on the positive and negative strands as shown in Supplementary Figure S9. The chromosomal distribution of known precursors largely matches with the distribution of the novel precursors as displayed in Supplementary Figure S10. In both cases, the least number of precursors can be found on chromosome Y. Chromosome 13, 18 and 21 harbor few known and novel precursors.

As for the number of motifs found in known and novel precursors with two annotated mature miRNAs, we found a slight enrichment of motifs in miRBase miRNAs (Supplementary Figure S11). A more fine-grained motif distribution is shown in Supplementary Figure S12.

Since miRNAs often occur in genomic clusters, we also searched genomic regions that are enriched by novel miRNAs. Supplementary Table S7 lists the positions of clusters when allowing a distance of at most 10 kb between the middle position of known or novel precursors. The largest cluster was composed of 46 known precursors and spanned 96 kb on chromosome 19. The largest cluster that contained both known and novel precursors was found on chromosome 14 and contained 42 known and 2 novel precursors and spanned 45 kb. In total 3969 clusters contained either known or novel precursors. Of these, 3423 clusters contained exclusively novel precursors. Further, 455 clusters contained both known and novel precursors and 91 exclusively known precursors. Supplementary Figure S13A and B shows the number of clusters with at least two or five precursors on each chromosome. Most clusters (394) with a minimum size of 2 could be found in chromosome 1. When focusing on clusters with at least five members, the numbers decreased to 154 clusters, 93 of which contained ex-

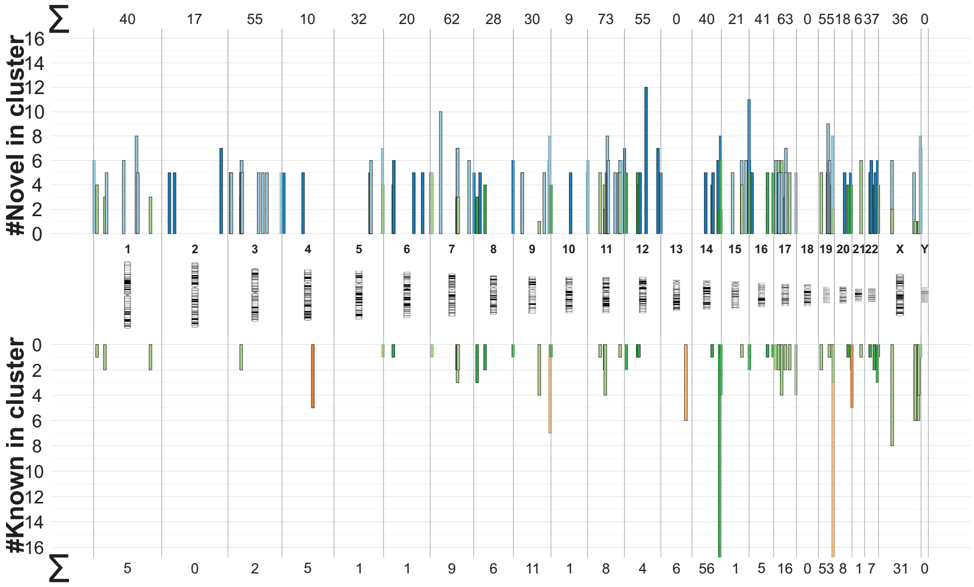


Figure 5. Distribution of the known and novel precursor clusters and their size on the human genome. Green clusters contain both novel and known precursors. Blue clusters contain only novel precursors and orange clusters contain only known precursors. The two known clusters on chromosome 14 and 19 (size 42 and 46) were trimmed for a better visualization. The sum of the number of novel or known precursors in all clusters of a chromosome with at least five members are shown on the top and bottom of the plot.

clusively novel precursors. Most clusters were observed on chromosome 11. Figure 5 shows the distribution of all clusters with five or more precursors over the human genome and demonstrates that many clusters contain both, known as well as novel precursors. The largest novel cluster with 12 precursors was found on chromosome 12.

To estimate how close our reported predictions might be to the coverage of the human miRNome, we performed predictions for different numbers of samples, each 10x randomly selected from our sample set. Supplementary Figure S14 shows the number of predictions according to the number of samples. We observe that the increase in number of predictions clearly exponentially diminishes with the number of samples. Since these predictions contain many false positives we expect the real part to be much smaller and the increase in predictions smaller as well. Therefore, we suggest that, at least for the tissues covered by our samples, we are close to the complete coverage of the human miRNome. We are aware and expect that the addition of samples of further tissue types or different conditions might add new candidates to our predicted set.

Expression analysis of miRNA candidates using custom microarrays

To provide further evidence that a relevant fraction of the aforementioned mature miRNAs is not only due to NGS bias or other artifacts such as RNA degradation, we built a custom human microarray. This array contains all miRBase v21 miRNAs, the miRNAs from the study by Londin *et al.* (30) and the top ranking miRNAs from the present study. The final microarray contained 11 866 miRNA candidates that have been measured each in 20 replicates (237 320 features per sample). For the microarray hybridization, we selected tissues from our Tissue Atlas (64) that contained the most miRNAs and added body fluids harboring likewise many miRNAs (65). The set of samples included a pool of PAXGene blood samples, a pool of plasma samples, lung tissue, brain tissue, kidney tissue, testis tissue, heart tissue and a reference pool from Agilent. Since degraded RNA is known to affect the miRNA patterns, we ensured high-quality of the used RNA samples. The RIN values of the different specimens ranged between 7.5 and 9. For the three sets of miRNAs the percentage of positive miRNAs in the hybridization experiments is presented in Figure 6A. For 56% of miRBase miRNAs, 55% of miRNAs by Londin *et al.* and 73% of miRNAs from the present study no positive signal in any sample was observed. On the other extreme, 11%, 17% and 8% were respectively positive in all experiments. The larger fraction of miRNAs not detected in any sample in the third set can be explained by the fact that many of the high abundant markers were previously already detected while we selected the candidates from the not yet discovered and likely much less abundant fraction. Still the results presented above can contain false positives (e.g. reagent contamination or positive signals induced by fragmented other RNAs) and false negatives (e.g. since other tissues or samples may harbor the miRNAs negative in the presently used samples or that are negative because of the limit of detection of microarrays). The same pattern as described can be recovered from the cluster analysis of all miRNAs from the three sets in Figure 6B. The lower part of this heat map shows that especially context sensitive miRNAs are observed among the set of miRNAs candidates only reported by miRMaster. In sum, the data

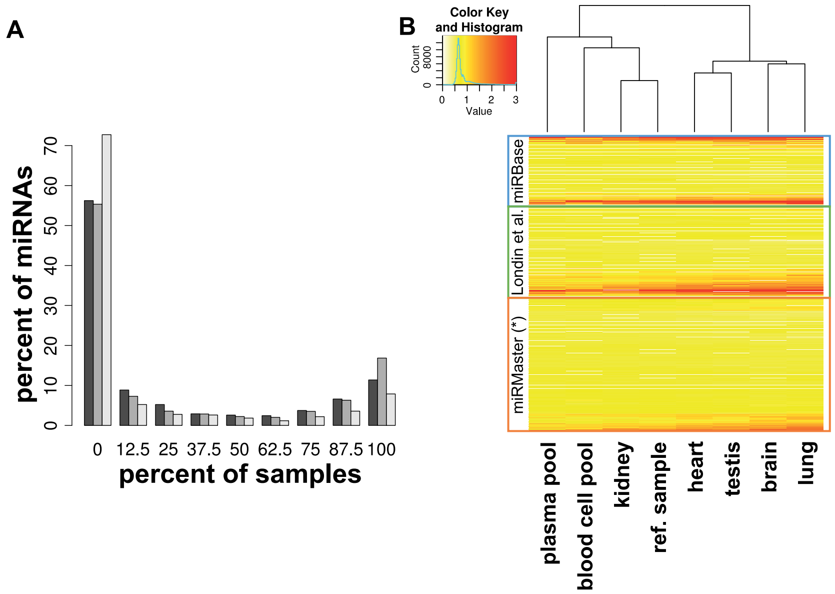


Figure 6. Expression of miRNA candidates on custom microarrays. (**A**) Distribution of the percentage of detected miRNAs in different samples. The colors correspond to the miRNAs of three studies: miRBase, dark gray; Londin *et al.*, medium grey; this study, light gray. (**B**) Heatmap of the logarithmized expression intensities of all miRNAs according to different tissues. For better visualization all expression values superior to 1000 were trimmed. The hierarchical clustering was performed with Euclidean distance.

strongly suggest that miRNAs exist which are currently not annotated in the miRBase. These miRNAs deserve further validation. All miRNAs from this analysis are contained in Supplementary Table S8.

CONCLUSIONS

The use of multiple web-based and standalone tools combined with different data formats makes the analysis of HT-seq miRNA data difficult, especially for wet-lab scientists. Therefore, we propose a web service that performs the most frequently requested applications directly from the raw FASTQ files. At the same time, experimental methods are advanced such that large-scale studies are feasible. Studies with many hundred or thousand samples are hard to be evaluated by current tools. Besides accuracy and specificity, runtime is among the most important criteria. Although miRMaster carries out a far greater number of analyses than other tools like miRDeep2, the running time of the miRMaster analysis was up to 20-fold faster. Of course, the precursor candidates predicted by miRMaster should in subsequent steps undergo a manual inspection and the selected ones be experimentally validated before calling them real miRNAs. A first validation step could be performed with our custom microarray followed by a more in depth validation of the detected interesting candidates using e.g. northern blotting. Applications such as target prediction, functional analysis and differential expression of known and novel miRNAs will in the future complete the portfolio of miRMaster.

ACCESSION NUMBERS

NGS samples are available on GEO under the following accession numbers: GSE64142, GSE53080, GSE49279, GSE45159 and GSE46579.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Best Ageing [306031] European Commission (in part); Saarland University; Siemens. Funding for open access charge: Internal funds of Saarland University.
Conflict of interest statement. None declared.

REFERENCES

1. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
2. Castellano,L. and Stebbing,J. (2013) Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. *Nucleic Acids Res.*, **41**, 3339–3351.
3. Chiang,H.R., Schoenfeld,L.W., Ruby,J.G., Auyeung,V.C., Spies,N., Baek,D., Johnston,W.K., Russ,C., Luo,S., Babiarz,J.E. *et al.* (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.*, **24**, 992–1009.
4. Jones-Rhoades,M.W. (2012) Conservation and divergence in plant microRNAs. *Plant Mol. Biol.*, **80**, 3–16.
5. Langenberger,D., Bartschat,S., Hertel,J., Hoffmann,S., Tafer,H. and Stadler,P.F. (2011) *Brazilian Symposium on Bioinformatics*. Springer, Vol. **6832**, pp. 1–9.

6. Meng,Y., Shao,C., Wang,H. and Chen,M. (2012) Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants. *RNA Biol.*, **9**, 249–253.
7. Tarver,J.E., Donoghue,P.C. and Peterson,K.J. (2012) Do miRNAs have a deep evolutionary history? *Bioessays*, **34**, 857–866.
8. Taylor,R.S., Tarver,J.E., Hiscock,S.J. and Donoghue,P.C. (2014) Evolutionary history of plant microRNAs. *Trends Plant Sci*, **19**, 175–182.
9. Wang,X. and Liu,X.S. (2011) Systematic curation of miRBase annotation using integrated small RNA high-throughput sequencing data for C. elegans and Drosophila. *Front. Genet.*, **2**, 25.
10. Fromm,B., Billipp,T., Peck,L.E., Johansen,M., Tarver,J.E., King,B.L., Newcomb,J.M., Sempere,L.F., Flatmark,K., Hovig,E. *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.*, **49**, 213–242.
11. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grasser,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
12. Hofmann,S., Huang,Y., Paulicka,P., Kappel,A., Katus,H.A., Keller,A., Meder,B., Stahler,C.F. and Gumbrecht,W. (2015) Double-stranded ligation assay for the rapid multiplex quantification of microRNAs. *Anal. Chem.*, **87**, 12104–12111.
13. Kappel,A., Backes,C., Huang,Y., Zafari,S., Leidinger,P., Meder,B., Schwarz,H., Gumbrecht,W., Meese,E., Staehler,C.F. *et al.* (2015) MicroRNA in vitro diagnostics using immunoassay analyzers. *Clin. Chem.*, **61**, 600–607.
14. Mestdagh,P., Hartmann,N., Baeriswyl,L., Andreasen,D., Bernard,N., Chen,C., Cheo,D., D'Andrade,P., DeMayo,M., Dennis,L. *et al.* (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.
15. Backes,C., Sedaghat-Hamedani,F., Frese,K., Hart,M., Ludwig,N., Meder,B., Meese,E. and Keller,A. (2016) Bias in High-Throughput Analysis of miRNAs and Implications for Biomarker Studies. *Anal. Chem.*, **88**, 2088–2095.
16. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
17. Hackenberg,M., Rodriguez-Ezpeleta,N. and Aransay,A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.
18. Rueda,A., Barturen,G., Lebron,R., Gomez-Martin,C., Alganza,A., Oliver,J.L. and Hackenberg,M. (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.*, **43**, W467–W473.
19. Guo,L., Yu,J., Liang,T. and Zou,Q. (2016) miR-isomiRExp: a web-server for the analysis of expression of miRNA at the miRNA/isomiR levels. *Sci. Rep.*, **6**, 23700.
20. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
21. Vlachos,I.S., Zagganas,K., Paraskevopoulou,M.D., Georgakilas,G., Karagkouni,D., Vergoulis,T., Dalamagas,T. and Hatzigeorgiou,A.G. (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.
22. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2004) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
23. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, doi:10.7554/eLife.05005.
24. Akhtar,M.M., Micolucci,L., Islam,M.S., Olivieri,F. and Procopio,A.D. (2016) Bioinformatic tools for microRNA dissection. *Nucleic Acids Res.*, **44**, 24–44.
25. Auyeung,V.C., Ulitsky,I., McGeary,S.E. and Bartel,D.P. (2013) Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell*, **152**, 844–858.
26. Fang,W. and Bartel,D.P. (2015) The menu of features that define primary microRNAs and enable de novo design of microRNA genes. *Mol. Cell*, **60**, 131–145.
27. Alarcon,C.R., Lee,H., Goodarzi,H., Halberg,N. and Tavazoie,S.F. (2015) N6-methyladenosine marks primary microRNAs for processing. *Nature*, **519**, 482–485.
28. Tatusova,T., Ciufo,S., Fedorov,B., O'Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.
29. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) MiRTargetLink–miRNAs, genes and interaction networks. *Int. J. Mol. Sci.*, **17**, 564.
30. Londin,E., Loher,P., Telonis,A.G., Quann,K., Clark,P., Jing,Y., Hatzimichael,E., Kirino,Y., Honda,S., Lally,M. *et al.* (2015) Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E1106–E1115.
31. Friedlander,M.R., Lizano,E., Houben,A.J., Bezdan,D., Banez-Coronel,M., Kudla,G., Mateu-Huertas,E., Kagerbauer,B., Gonzalez,J., Chen,K.C. *et al.* (2014) Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.*, **15**, R57.
32. Jha,A., Panzade,G., Pandey,R. and Shankar,R. (2015) A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Res.*, **43**, 8713–8724.
33. Leidinger,P., Backes,C., Deutscher,S., Schmitt,K., Mueller,S.C., Frese,K., Haas,J., Ruprecht,K., Paul,F., Stahler,C. *et al.* (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.*, **14**, R78.
34. Keller,A., Leidinger,P., Vogel,B., Backes,C., ElSharawy,A., Galata,V., Müller,S., Marquart,S., Schrauder,M., Strick,R. *et al.* (2014) miRNAs can be generally associated with human pathologies as exemplified for miR-144*. *BMC Med.*, **12**, 224.
35. Backes,C., Leidinger,P., Altmann,G., Wuerstle,M., Meder,B., Galata,V., Mueller,S.C., Sickert,D., Stahler,C., Meese,E. *et al.* (2015) Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. *Anal. Chem.*, **87**, 8910–8916.
36. Roth,P., Keller,A., Hoheisel,J.D., Codo,P., Bauer,A.S., Backes,C., Leidinger,P., Meese,E., Thiel,E., Korfel,A. *et al.* (2015) Differentially regulated miRNAs as prognostic biomarkers in the blood of primary CNS lymphoma patients. *Eur. J. Cancer*, **51**, 382–390.
37. Keller,A., Leidinger,P., Meese,E., Haas,J., Backes,C., Rasche,L., Behrens,J.R., Pfuhl,C., Wakonig,K., Giess,R.M. *et al.* (2015) Next-generation sequencing identifies altered whole blood microRNAs in neuromyelitis optica spectrum disorder which may permit discrimination from multiple sclerosis. *J. Neuroinflammation*, **12**, 196.
38. Schwarz,E.C., Backes,C., Knorck,A., Ludwig,N., Leidinger,P., Hoxha,C., Schwar,G., Grossmann,T., Muller,S.C., Hart,M. *et al.* (2016) Deep characterization of blood cell miRNomes by NGS. *Cell. Mol. Life Sci.*, **73**, 3169–3181.
39. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement*, **12**, 565–576.
40. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
41. Sacar,M.D., Hamzeiy,H. and Allmer,J. (2013) Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? *J. Integr. Bioinform.*, **10**, 215.
42. Chou,C.H., Chang,N.W., Shrestha,S., Hsu,S.D., Lin,Y.L., Lee,W.H., Yang,C.D., Hong,H.C., Wei,T.Y., Tu,S.J. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.
43. Kim,V.N., Han,J. and Siomi,M.C. (2009) Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **10**, 126–139.
44. Xue,C., Li,F., He,T., Liu,G.P., Li,Y. and Zhang,X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.
45. Babiarz,J.E., Ruby,J.G., Wang,Y., Bartel,D.P. and Blelloch,R. (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev.*, **22**, 2773–2785.

46. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
47. Ng,K.L. and Mishra,S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321–1330.
48. Batuwita,R. and Palade,V. (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, **25**, 989–995.
49. Lertampaiporn,S., Thammarongtham,C., Nukoolkit,C., Kaewkamnerdpong,B. and Ruengjitchatchawalya,M. (2013) Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. *Nucleic Acids Res.*, **41**, e21.
50. Lee,M.T. and Kim,J. (2008) Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS Comput. Biol.*, **4**, e1000150.
51. Zhang,B.H., Pan,X.P., Cox,S.B., Cobb,G.P. and Anderson,T.A. (2006) Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci.*, **63**, 246–254.
52. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Methodol.*, **57**, 289–300.
53. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
54. Kim,J., Levy,E., Ferbrache,A., Stepanowsky,P., Farcas,C., Wang,S., Brunner,S., Bath,T., Wu,Y. and Ohno-Machado,L. (2014) MAGI: a Node.js web service for fast microRNA-Seq analysis in a GPU infrastructure. *Bioinformatics*, **30**, 2826–2827.
55. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
56. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
57. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
58. Zhang,P., Si,X., Skogerbo,G., Wang,J., Cui,D., Li,Y., Sun,X., Liu,L., Sun,B., Chen,R. *et al.* (2014) piRBase: a web resource assisting piRNA functional study. *Database (Oxford)*, **2014**, bau110.
59. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
60. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
61. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
62. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
63. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
64. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stahler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
65. Fehlmann,T., Ludwig,N., Backes,C., Meese,E. and Keller,A. (2016) Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol.*, **13**, 1084–1088.

Published online 19 April 2021 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* ***W397–W408***
https://doi.org/10.1093/nar/gkab268

miRMaster 2.0: multi-species non-coding RNA sequencing analyses at scale

Tobias Fehlmann[1], Fabian Kern[1], Omar Laham[1], Christina Backes[1], Jeffrey Solomon[1], Pascal Hirsch[1], Carsten Volz[2], Rolf Müller[2] and Andreas Keller[1,3,*]

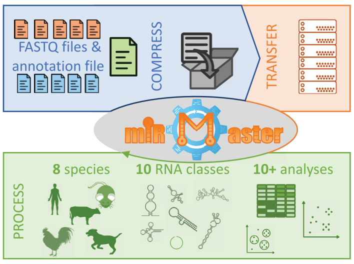
[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Department of Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmacy, Saarland University, Campus E8 1, 66123 Saarbrücken, Germany and [3]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA

Received February 24, 2021; Revised March 21, 2021; Editorial Decision March 30, 2021; Accepted April 15, 2021

ABSTRACT

Analyzing all features of small non-coding RNA sequencing data can be demanding and challenging. To facilitate this process, we developed miRMaster. After the analysis of over 125 000 human samples and 1.5 trillion human small RNA reads over 4 years, we present miRMaster 2 with a wide range of updates and new features. We extended our reference data sets so that miRMaster 2 now supports the analysis of eight species (e.g. human, mouse, chicken, dog, cow) and 10 non-coding RNA classes (e.g. microRNAs, piRNAs, tRNAs, rRNAs, circRNAs). We also incorporated new downstream analysis modules such as batch effect analysis or sample embeddings using UMAP, and updated annotation data bases included by default (miRBase, Ensembl, GtRNAdb). To accommodate the increasing popularity of single cell small-RNA sequencing data, we incorporated a module for unique molecular identifier (UMI) processing. Further, the output tables and graphics have been improved based on user feedback and new output formats that emerged in the community are now supported (e.g. miRGFF3). Finally, we integrated differential expression analysis with the miRNA enrichment analysis tool miEAA. miRMaster is freely available at https://www.ccb.uni-saarland.de/mirmaster2.

GRAPHICAL ABSTRACT



INTRODUCTION

The reliable analysis of small non-coding RNA (sncRNAs) sequencing data can be challenging, time consuming and varies in many aspects. This includes the primary processing of sequencing data for quality control but also downstream analyses. Besides tRNAs and snoRNAs, microRNAs are sncRNAs that are extensively studied already for over two decades. A comprehensive summary on the state-of-the art in microRNA biology has been published by Bartel in 2018 (1). In addition to the canonical biology of microRNAs, more and more non-canonical aspects of miRNA biology are becoming obvious (2), calling for a broad variety of analysis aspects. It is thus not surprising, that many miRNA analysis tools, online and standalone, are available. The current release of Aviator (https://www.ccb.uni-saarland.de/aviator), a tool that aims to provide accessibility statistics of all web servers and data bases in life sciences, lists 322 web-based resources for microRNAs, of which 235 are currently working. By developing miRMaster (3,4), we provide a tool with a strong focus on the analysis of all aspects of miRNAs described in a systematic manner. While the tool became stepwise broader in its functionality aspect to cover other sncRNA classes, microRNAs are still its anchor point. Based on the original version of miRMaster, 1500 runs have been completed, over

[*]To whom correspondence should be addressed. Tel: +49 681 302 68610; Email: andreas.keller@ccb.uni-saarland.de

125 000 human samples have been analyzed, and 1.5 trillion human small RNA reads were processed over four years. To make further use of the uploaded data, we ask miRMaster users whether we can re-analyze the aggregated sequencing reads as further feedback for the development of our tool and as comparison to our small RNA research projects.

As mentioned before, several other web servers and webservices for analyzing miRNA sequencing data exist, overlapping partially with miRMaster's functionality. We thus want to put our tool in the context of others and recent developments in the field. A broad tool meta review has been published by Chen *et al.*, which covers 95 review papers and about 1000 miRNA bioinformatics tools (5). The variety of tools reaches from rather specialized tools, e.g. for the detection of isomiRs or miRNA editing (6) to very broad analysis pipelines. Among the tools with broader analysis functionality with a particular focus on miRNAs, we want to mention CBS-miRSeq (7), miRquant (8), sRNAbench/sRNAtoolbox (9), Chimira (10), mirPRo (11), miRge (12) and CPSS2 (13). Among those, sRNAtoolbox was updated most recently with improvements made to e.g. the batch processing, library preparation protocols and differential expression analysis. Moreover, SPAR is a small RNA-seq portal for analysis of sequencing experiments (14). GLASSgo for example facilitates automated detection of sRNA homologs (15). Also, for viruses, sRNA analysis tools have been developed such as MISIS-2, a tool for the in-depth analysis of sRNAs and representation of consensus master genomes in viral quasispecies (16). With respect to non-canonical miRNA biology, such as miRNA sponges (17), also specialized software tools have been made available (18). Another aspect is the analysis of miRNA editing and chemical modifications. Likewise, for this task, specialized tools are available such as Prost! (19), DeAnniso (20) and others. Similarly, for tRNAs and other ncRNA classes, chemical modifications exist (21) that further complicate the analysis of ncRNA data sets.

The number of available tools, web-based and standalone, reflects the continuous interest of the research community in non-coding RNAs. At the same time, it pinpoints that a more integrative analysis of different ncRNA classes is required. Recently, a 'changing of the guards' model has been proposed, where microRNA levels decreased but small transfer RNA fragments increased in blood of patients (22).

Although a direct comparison in terms of scope, functionality, ease of use and other parameters is challenging and partially subjective, we aimed to provide at least an overview on a selection of commonly used broader analysis tools that are available as web-service. We thus evaluated the number of analyses, the number of supported organisms, the number of supported ncRNA classes and other parameters for 10 tools and present the result sorted by the publication date (Figure 1). The analysis reveals an expected pattern, the more recent tools have a broader scope of functionality as compared to the early tools. The recently updated sRNABench for example excels in basically all categories, e.g. offers more organisms as compared to miRMaster 2. miRMaster 2 offers in contrast more output options and covers more ncRNA classes. Figure 1 allows to compare the functionality of the 10 selected tools and supports users in their decision to select one tool.

MATERIALS AND METHODS

Reference data bases

miRMaster relies on public annotation data sets of several well-known and widely used reference data bases. These include for the different RNAs miRBase (version 22.1 (23)), Ensembl ncRNA (version 100 (24)), RNACentral (for piRNAs) (version 15 (25)), GtRNAdb (version 18.1 (26)), circBase (accessed 25.10.20 (27)), NONCODE (version 5 (28)) and NCBI RefSeq for the reference genomes as well as viruses and bacteria.

Supported sequencing protocols

miRMaster 2 directly supports the most common sequencing protocols. This includes Illumina Truseq, Bioo Scientific Nextflex, MGISeq and Diagenodes D-Plex and CATS technology.

Implementation and graphical representation of the web service

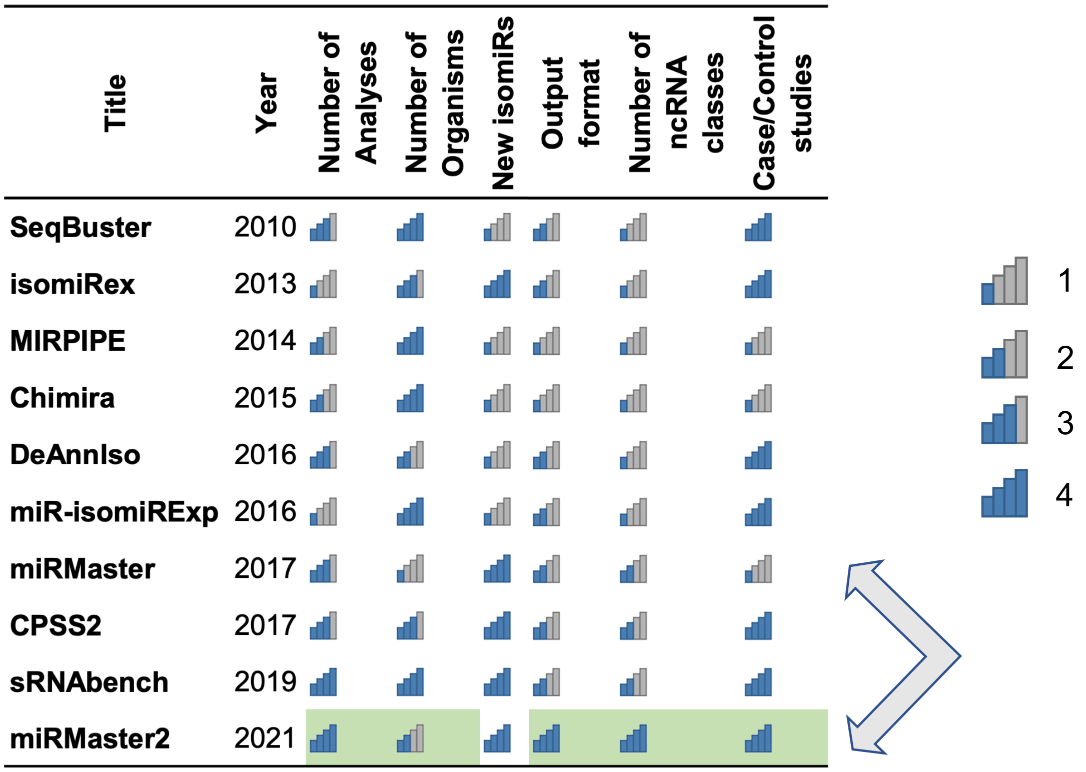
The miRMaster 2 web service was implemented using Python 3.7.6 with Django 2.2.10, Postgres 11.1 and Redis 5.0. All services are encapsulated in docker containers and bundled with docker compose. The job queue is handled with Celery 4.4.7 and Redis, and the jobs are executed via Snakemake 5.31.1 (29). The frontend was styled with Bootstrap 4.5.3 and the interactive features are based on the Angular JS 1.5.11 and jQuery 3.4.1 libraries. Plots are rendered with Highcharts 8.2.2 and Clustergrammer-GL 0.22.0 (30). RNA secondary structures are rendered with fornac 1.1.10 and interactive tables with DataTables 1.10.23.

Compression and quality control data

miRMaster accepts raw FASTQ and gzip compressed FASTQ files as input. At first and before the data is sent to our server, we perform three pre-processing steps encompassing adapter trimming, quality filtering and read collapsing via JavaScript on the upload page. The collapsed reads are transmitted to the servers as soon as possible but in chunks of ∼16MB, thereby ensuring a low RAM consumption. The quality scores of the reads are collected on the user side and only aggregated metrics are sent to the server. Multiple annotations can be provided for the uploaded samples, which can then be highlighted in the resulting reports. In particular, a group annotation for differential expression downstream can be selected, in which case miRMaster will perform additional analyses.

Data pre-processing

While most of the analysis parameters in the pre-processing view are conveniently set to reasonable default values, the expert mode allows full control and maximal flexibility. In the pre-processing view, the 5′ barcode length, an adapter barcode length and the length of the unique molecular identifier (UMI) can be determined. Further, after activation of the expert mode, minimum read length and maximum adapter edit distance can be modified as well as the minimum read/adapter overlap. Finally, leading and trailing



Title	Year	Number of Analyses	Number of Organisms	New isomiRs	Output format	Number of ncRNA classes	Case/Control studies
SeqBuster	2010	3	4	1	2	1	4
isomiRex	2013	1	3	4	2	1	4
MIRPIPE	2014	2	4	1	1	1	1
Chimira	2015	2	4	1	1	1	1
DeAnnIso	2016	3	2	1	2	1	4
miR-isomiRExp	2016	1	4	1	2	1	4
miRMaster	2017	3	1	4	2	2	1
CPSS2	2017	3	3	4	2	2	4
sRNAbench	2019	4	4	4	2	2	4
miRMaster2	2021	4	2	4	4	4	4

	1	2	3	4
Number of Analyses	one analysis	up to three analyses	up to 6 analyses	over 6 analyses
Number of Organisms	one organism	up to 10 organisms	up to 100 organisms	over 100 organisms
New isomiRs	no	-	-	yes
Output format	tables	tables and images	-	tables/interactive images
Number of ncRNA classes	no	up to 8 other classes	-	over 8 other classes
Case/Control studies	no	-	-	yes

Figure 1. Tool comparison. Comparison of features provided by tools analyzing sncRNA-seq data. Improvements of miRMaster 2 in comparison to its original release are marked in green.

N's can be trimmed, or reads containing N's can be fully omitted. For quality trimming, the sliding window size and quality can be selected. To maximize the processing capabilities of the user's CPU, the number of threads that can be used on the client side for the pre-processing can be configured.

Mapping

The read mapping process in miRMaster 2 is carried out using Bowtie 1.2.3 (31) as the standard option. For the update we added STAR 2.7.5a (32) as alternative mapper. In the standard mode, up to five hits in the reference genome are allowed for each read, where the mapping seed length is set to 18 nt and no mismatches in the seed region are allowed. In the expert mode, the user is free to change these parameters. For quantifying miRNAs, reads are mapped against miRNA precursors while allowing per default one mismatch. The resulting mappings are then filtered to count only reads mapping to the annotated miRNAs with at most two nucleotides differences at the 5′ end and five nucleotides at the 3′ end. For isomiR quantification, the mapping is performed with one additional mismatch and then subsequently filtered, such that non-templated nucleotide additions are not counted as mismatches. Other ncRNAs are per default quantified without any mismatches and all multimapping reads with the lowest number of mismatches are considered. To generate the miRGFF3 isoform format, mirtop 0.4.23 (33) is run on the aligned files and subsequently filtered to accommodate the user selected number of allowed mismatches.

Detection filtering

Before applying analysis methods to the expression matrices of the different ncRNA types, these matrices are filtered per default, such that only those RNAs that are expressed in at least 50% of all the samples are kept, or in case a differential expression annotation variable is provided (e.g. Diagnosis), in at least 50% of the samples of one of the variable levels (e.g. Dementia or Control). For this filtering procedure, only RNAs that are expressed with at least three reads are considered detected. In addition to normalizing the reads by the sequencing depth (RPM) we $\log_2$ transform the expression data and add a pseudo count of one.

Embedding

miRMaster 2 is equipped with two common dimension reduction approaches, namely Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). The user can first select the embedding from a drop down followed by the response variable. This response variable is extracted from the annotation file and the data points are colored according to the respective grouping. As representation, 2D-scatter plots are provided to the user.

Clustering

Hierarchical clustering with Euclidean distance and complete linkage is performed on the sample Spearman correlation, which is determined based on the reads per million (RPM) normalized expression matrix for each ncRNA type separately, as well as for all sncRNAs. In addition, hierarchical clustering is also performed on the RPM $\log_2$ normalized expression matrix for each ncRNA type, as well as on subsets of the top RNAs with the largest variance.

Batch effect analysis

To detect the influence of technical batches or attribute variance to biologically relevant parameters, a Principal Variance Component Analysis (PVCA) is performed. PVCA combines the strengths of two data analysis techniques, principal component analysis (PCA), which reduces the feature dimensions while maintaining the largest fraction of the variability in the data, and variance components analysis (VCA), which fits a mixed linear model using factors of interest as random effects. In more detail, all variables provided in the annotation file are fit as random effects including two-way interaction terms in the mixed model. Thereby, principal components obtained from the original data expression matrix are selected. As a result, the proportion of variance that is attributed to the variables from the annotation file is reported.

Differential gene expression analysis

The groupings for differential expression analysis are extracted from the annotation file provided by the user. If more than two groups are given, all pair-wise differential expression analyses are performed automatically. miRMaster first computes whether the features are normally distributed by applying the Shapiro Wilk test. As hypothesis test for assessing the degree of differential expression, *t*-test (for normally distributed data) and Wilcoxon Mann–Whitney test otherwise, are calculated. For multiple groups, also analysis of variance (ANOVA) as well as the non-parametric Kruskal–Wallis test are computed. The *P*-values are adjusted for multiple testing by controlling the false discovery rate using the Benjamini–Hochberg procedure. Further measures representing effect sizes are fold changes, the area under the receiver operator characteristics curve (AUC value) and Cohen's *d*. In addition to tabular output, volcano plots ($\log_2$ fold change versus negative decade logarithm of *P*-values) are displayed and boxplots per RNA are shown.

miRNA prediction

The prediction of new miRNAs follows the same principles as described in miRDeep (34) and our previous publications (3,4). Similarly, the expert mode allows maximal flexibility. As a first step, after the reads have been mapped to the genome, miRNA precursor candidates are determined. To this end local maximum read stacks, which are assumed to stem from potential miRNAs, are searched in downstream windows of per default 70 nucleotides and two precursors are excised from each stack. For this step, the required minimum read stack can be increased in order to improve the specificity of miRNA precursor predictions. Also, minimal and maximal length of the mature miRNA(s) can be set. Moreover, the mapping fraction consistent with Dicer processing can be increased or decreased. miRMaster groups new miRNA candidates into several categories, depending on their overlap with known miRNAs or known miRNA precursors. Precursor miRNA candidates belonging to the 'novel' category are not overlapping any known miRNAs and none of their annotated miRNAs have a similarity to any known miRNA of the same species. Details on the different categories are provided in the software tutorial page. To further rank and prioritize the predicted precursors, NovoMiRank (35) is subsequently applied and the scores are presented in the results table.

API to miEAA

Following our ambition to develop a fully integrated knowledge base on miRNAs (36), we started to integrate our tools such as miEAA (37,38) and miRSwitch (39) with APIs. For miRMaster we continued this process in the reverse direction. miRNAs can be sorted by their expression level, or if case–control studies are evaluated, also with respect to their differential expression. From the miRMaster results page, the miEAA APIs for over-/underrepresentation and miRNA set enrichment analyses can be queried such that functional pathway enrichment analyses are feasible within minutes. Moreover, miRNAs are linked to target genes and target gene networks using miRTargetLink 2 (40), when applicable.

RESULTS

In this section, we aim for a complete and self-containing description of miRMaster 2, partially sketching features that have already been available in the original version such as the miRNA prediction module. The improvements are highlighted, summarized in the conclusion and also marked in green in Figure 1.

Data input

miRMaster relies on FASTQ files (optionally gzip compressed) as main input. Additionally, an annotation file can be provided by users to facilitate downstream analyses. While the annotation file is typically small, represented by few kilobytes of data, the FASTQ files can encompass several gigabytes per sample. Therefore, the data transfer for large studies could become a time-consuming task. We thus implemented an algorithm, exploiting the fact that miRNA-seq libraries are often of lower read complexity, that compresses the data typically by over 90%. This happens at the client side and only the compressed data are then transferred. Similarly, selected quality statistics are computed at the client side and only the relevant compressed information is transmitted. The server-side processing of the data starts immediately, i.e., while the data are transferred the processing is already initiated. We optimized these procedures in a way that studies with several hundred samples can be processed by miRMaster. The time-consuming analyses carried out at the server side afterwards are mostly implemented in efficient C++, such that even large-scale studies exceeding hundreds of samples are processed within a few hours.

Supported organisms and RNA classes

While the original version of miRMaster was centered around the analysis of human microRNA data, supporting only a few other small RNA data analyses, we now provide full support for several other common organisms and more ncRNA classes. Importantly, the analysis of the RNA classes is not only restricted to small non-coding RNAs but also to longer RNAs such as circRNAs. With respect to the organisms, miRMaster 2 supports, besides *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Monodelphis domestica*, *Macaca mulatta*, *Gallus gallus*, *Bos taurus* and *Canis familiaris*.

Pre-processing functionality and mapping

In the pre-processing step, adapters are trimmed and only sequences exceeding the selected minimum length are extracted. Furthermore the *GC* content is calculated. The results of this step are available as table and displayed as boxplots (Figure 2A). If the user provided several groups, as for instance cases and controls, or grades of severity for a disease, the information is displayed per group. In addition to the aggregated statistics, detailed per-sample statistics also are available. All tables can be downloaded in excel and csv format. All graphics are available as jpg, pdf, png and svg files. The underlying data for each graphic panel can also be downloaded in case users want to make figures on their own.

The second analysis step is genome and non-coding RNA mapping. First, the mapping to the reference genome is performed using the user specified input parameters and mapper. Mapping to the following 10 RNA classes is also performed: microRNA (Figure 2B), tRNA, piRNA, rRNA, scaRNA, lncRNA, snoRNA, snRNA, miscRNA and circRNA (Figure 2C). Additionally, mapping against viruses and bacteria from RefSeq is carried out for each sample, based on reads that did not map against the reference genome. As for the adapter trimming, all available information can be downloaded in excel and csv format and for each sample detailed mapping statistics are presented.

Sample embedding, clustering, batch effect analysis

After completing the pre-processing and mapping, different aggregated analyses on the sample and RNA class level are carried out. First, an embedding using UMAP or alternatively PCA is available. The embedded graphics as 2D scatter plot can be colored with respect to arbitrary input variables extracted from the annotation file. For each RNA class, a distinct embedding is available (miRNA Figure 2D, tRNA Figure 2E). This allows users for example to visually compare whether for one RNA class a better clustering with respect to a disease phenotype is observed as compared to another.

Next, a sample correlation analysis is performed. Here, the correlation of the RNA expression between all pairs of samples is computed and shown as heatmap. The ordering of rows and columns can be modified and colored representations on top of the heatmaps show for each provided variable a color code (Figure 2F). Per default, hierarchical clustering with Euclidean distance and complete linkage is shown. In addition to the sample-to-sample correlation clustering, also the clustering of the expression values for each of the 10 RNA classes is computed, with the possibility to focus on the subset of RNAs with the largest variance. For both, rows (features) and columns (samples), clusters are defined and automatically adjusted. If a cluster is selected, the distribution of representatives within the cluster is displayed and *P*-values for enrichments are provided.

As last consideration of this functionality aspect, miRMaster 2 estimates the proportion of variance that can be attributed to variables provided in the annotation file. This can be biologically relevant metadata (again, case or control, or different severity grade of a disease, sex and many others) or it can be technical batches (e.g. the information which samples have been sequenced together or come from the same site in a multi-centric study). To this end, miRMaster 2 performs a PVCA, that combines aspects of principal component analysis as well as variance components analysis. The results are provided as bar charts, again for each of the RNA classes separately (Figure 2G for miRNA and Figure 2H for tRNAs). While miRMaster 2 highlights potential experimental batches it currently does not provide function-

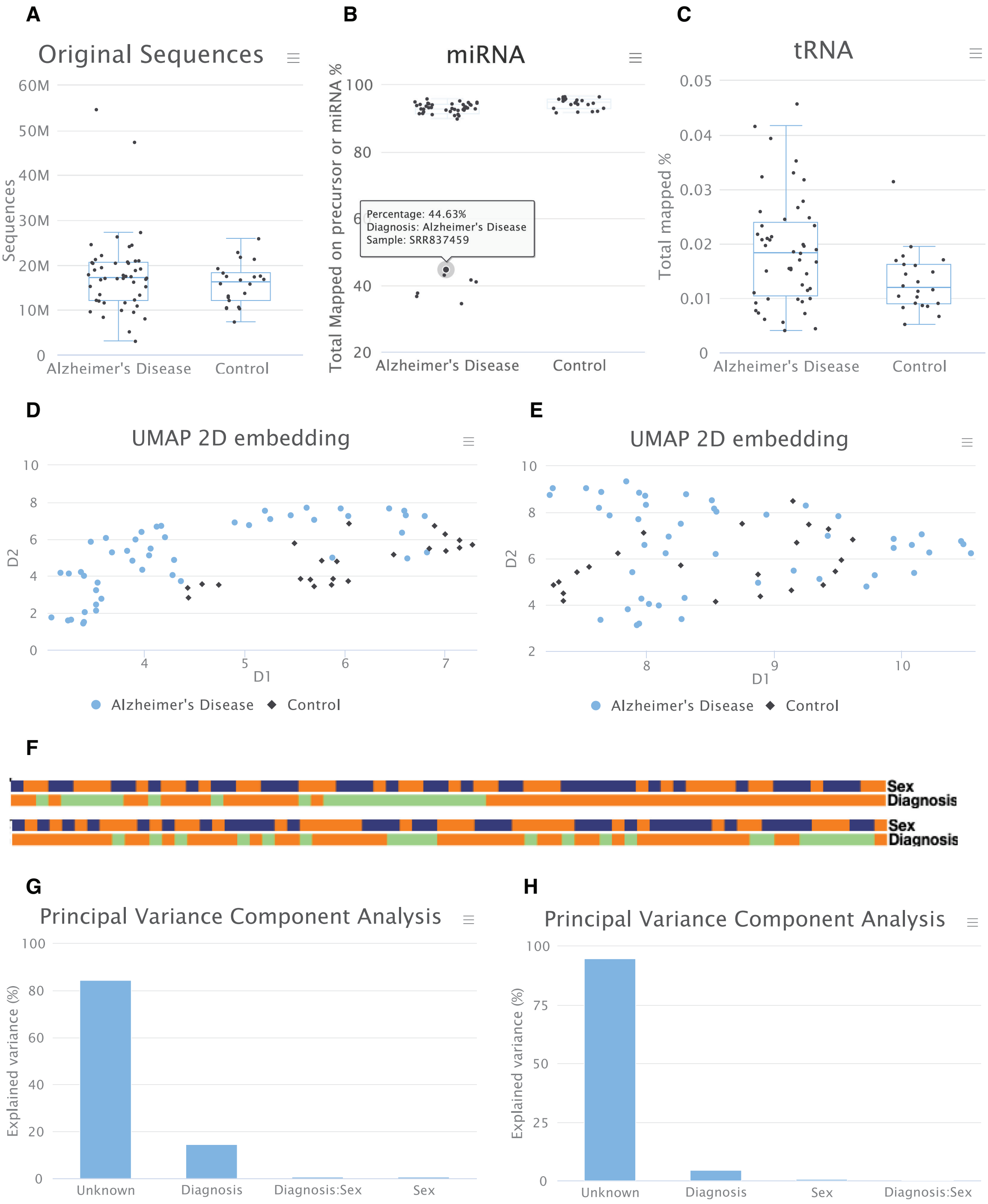


Figure 2. Selected result for the data pre-processing. The presented data are taken from the online demo data set on Alzheimer's Disease (AD). (**A**) Number of reads in the data set. Each dot represents a single sample. No difference between AD and controls exists. (**B**) Mapping to microRNAs. One point is highlighted. This feature can be used to identify outliers. (**C**) Mapping to tRNAs. In the overall distribution we observe here differences between the two classes. (**D**) Embedding of the samples using UMAP, colored by the disease phenotype using miRNAs. A clustering in the two groups can be recognized in this embedding. (**E**) The same embedding for tRNAs. In the case of this RNA class, no clear clustering is present. (**F**) Color-coded clustering for the sex and disease phenotype for microRNAs (top) and tRNAs (bottom). The two classes don't show clustering with respect to sex but for microRNAs a clustering of AD samples can be observed. (**G**) Results of the PVCA for miRNAs. Around 15% of the total variance can be explained by the disease phenotype. (**H**) The same results for the tRNAs. Here, a lower percentage of variance is explained by the disease phenotype.

ality to correct for batch effects. If experimental batches exist, a batch correction by the user is therefore recommended.

Differential expression, downstream analysis and APIs to other tools

The previous analyses are based on sets of non-coding RNAs without focusing on single feature expression or differential expression, one of the key features of miRMaster 2. First, the expression of each miRNA is computed and shown in an aggregated and per-sample manner. The miRNAs can be sorted according to expression levels and pathway analyses using miEAA (all miRNAs) or miRTargetLink (single miRNA-gene interactions) are available. The results table can be downloaded as excel file or in csv format. If an annotation file was specified and a differential expression variable defined, volcano plots are generated. These show the negative decade logarithm of the *P*-value (adjusted Wilcoxon Mann–Whitney test) versus the $\log_2$ fold change (Figure 3A). From the interactive results table, miRNAs can be selected and boxplots detailing the expression of each sample are computed (Figure 3B/C). This table contains the raw and adjusted *P*-values of the Wilcoxon Mann–Whitney test as well as the results for a Shapiro Wilk normality test and *t*-test. If the data are normally distributed, t-test *P*-values can be used instead of the Wilcoxon Mann–Whitney test *P*-values. In addition, the table lists ANOVA and Kruskal–Wallis test *P*-values in case of more than two categories need to be compared per variable. As measure for the effect size, in addition to the fold change, the area under the receiver operator characteristics curve (AUC) is computed, as well as Cohen's d. For the de-regulated miRNAs, miRNA set enrichment analysis as well as over-representation analysis is facilitated through miEAA. The same metrics on differential expression are computed for each feature and all other RNA classes. In the case of other ncRNAs, however, no pathway enrichment analysis is available currently. This limitation might however be solved in the future by adding enrichment analyses for target genes or by integrating future data bases on pathways for other ncRNA classes that are developed.

A key question for researchers is to select the most valid de-regulated candidates and to exclude likely false positives. From our experience, several factors contribute to the success in validating de-regulated miRNAs (41). The miRNA has to show sufficient expression, relevant effect sizes between cases and control and at best a statistically significant difference. To allow users filtering for the best candidates, the results table offers the option to add several filter criteria that are connected by a logical 'and'. Authors can filter for those miRNAs present with at least 1 RPM, having an effect size of at least 0.7 and a *P*-value <0.05. By adapting the parameters, users can balance for rather specific results or rather sensitive results, depending on the underlying biological question.

Remarkably, miRMaster 2 supports the analysis of multiple comparisons at the same time. If the annotation column that is selected by the user has for example four groups, all pair-wise comparisons ($4 \times 3/2 = 6$) are carried out and presented to the user. Currently, however, only one annotation column can be used at a time for an analysis to avoid too many computations and complicated result representations.

The last analysis module comprises the prediction of new miRNAs. Again, the same information on expression and de-regulation as for the 10 RNA classes is calculated and presented to the user. Here, novel precursor miRNAs found in the data can be selected. An example is a precursor miRNA where the 5′ mature form is annotated while the 3′ that is expressed in the user's data is not yet annotated. For each miRNA candidate, the secondary structure and free energy is computed and presented (Figure 3D). Finally, the expression of reads on the 5′ and 3′ mature form (Figure 3E) is provided as bar-plot along the precursor. Here, users can verify the read stacks manually and in principle observe potential 3′ heterogeneity. Next, all the mapping reads are presented, containing potential isomiRs or other RNA fragments (Figure 3F). This module offers full-download capabilities as enumerated for the other analysis modules.

We clearly consider the respective potentially new miRNAs as candidates only. It is frequently hard to distinguish between molecules of the different ncRNA classes (e.g. miRNAs that are actually tRNA fragments) or to exclude artifacts. The candidates call for an in-depth experimental validation before they should be considered as miRNAs (42). The validation rate is tightly coupled with the quality of the pre-selected precursors. In our previous study we reached a validation of almost 20%.

In the following sections we provide three use cases, first the analysis of a sncRNA-seq data set from *Mus musculus*, second, human sncRNA data from dementia patients and finally the analysis of single-cell small non-coding RNA data using unique molecular identifiers. All use cases are available on the miRMaster homepage.

Use Case (1): *Mus musculus* sncRNA atlas

To demonstrate the analysis scope of miRMaster for a non-human species, we analyzed sncRNA sequencing data from *M. musculus* (43). Here, for 11 organs and up to 14 replicates of both sexes, sncRNA data using Illumina sequencing have been analyzed. We started miRMaster with the 272 FASTQ files downloaded from SRA, used the tissue and mouse ID as annotation and configured the sex as differential expression variable. The complete results were available after 4 hours. The pre-processing page highlights a large span of sequencing depth, going from a few thousands up to 25 million reads, with a median of 1.9 million reads for the male samples and 5.1 million reads for the female samples. The mapping statistics show overall similar patterns between male and female samples for all considered RNA classes. Interestingly, the miRNA mapping statistics show the highest variability, going from less than 5% mapped reads for some samples, up to over 90% of mapped reads. This variability can mainly be attributed to the RNA composition of the different tissues, as shown by the batch analysis, where 73% of the observed variance can be explained by the tissue variable (Figure 4A). We find that for all other RNA classes, although varying in the proportion of explained variance, the tissue is the strongest factor, except for circRNAs, where most of the observed variance is explained by the sex (Figure 4B). These results are reflected by the

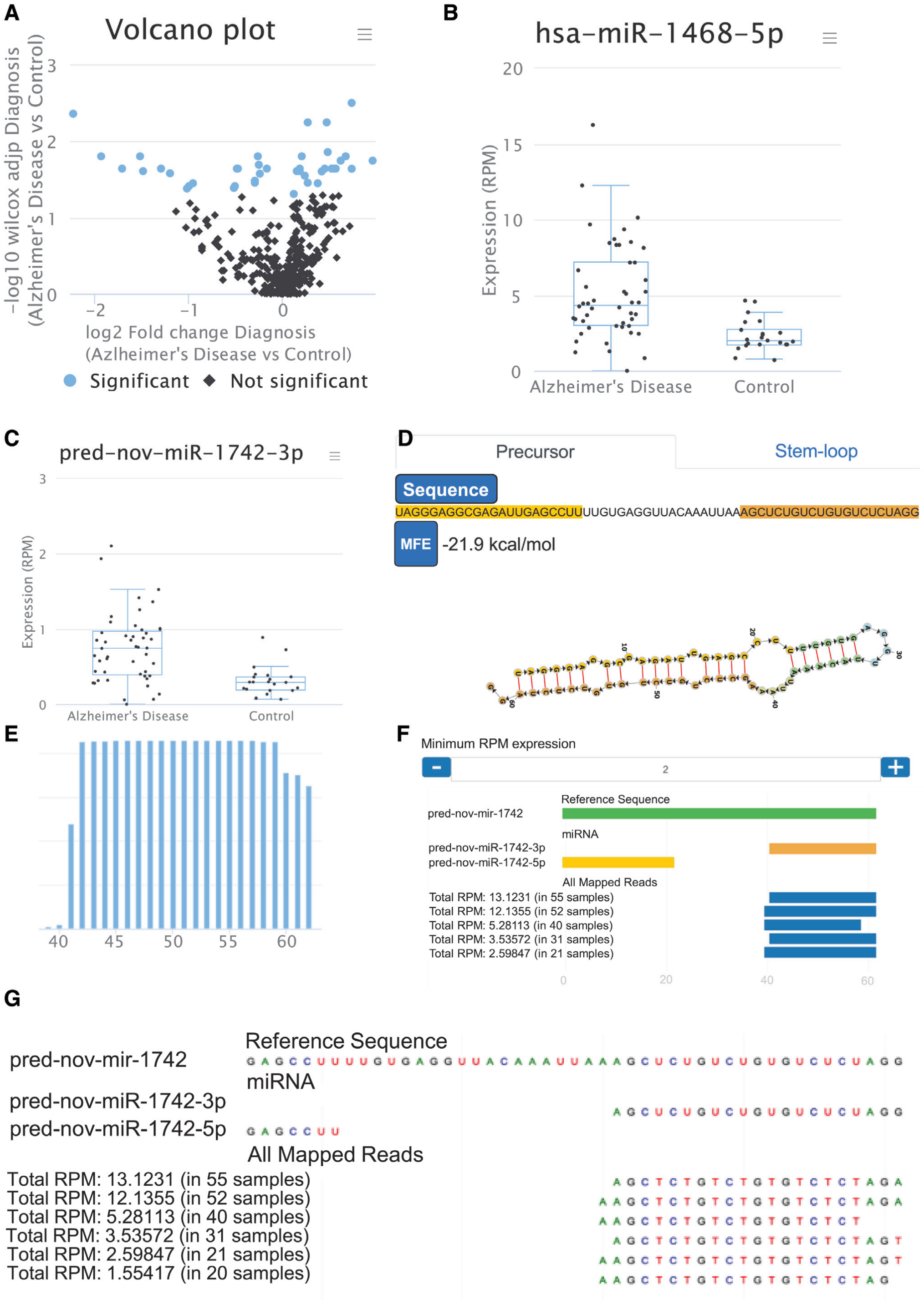


Figure 3. Downstream analyses for AD and control samples. (**A**) Volcano plot displaying each microRNA as a dot. Colored dots are statistically significant (adjusted P-value < 0.05). (**B**) For one significant marker (hsa-miR-1468–5p) the data are presented as boxplots. Again, single samples can be highlighted by moving the mouse over. (**C**) Box-plot for a novel miRNA candidate. (**D**) Precursor structure of this novel miRNA candidate and the minimum free energy. Users can switch between the representation of the precursor and the stem-loop. (**E**) Distribution of reads on the mature -3p miRNA of the same candidate precursor. Towards the end of the mature miRNA, the –3p heterogeneity that is typical for miRNAs can be recognized. (**F**) Representation of isoforms. The green bar denotes the precursor, the yellow bar the –5p mature form, the orange bar the –3p form. Each blue bar shows an isoform. The number of reads supporting the isoform can be dynamically adjusted by the user (here, at least 2 RPM are required). (**G**) If users zoom in the representation, the single base resolution per isoform is displayed (in this example, 1 RPM coverage is sufficient).

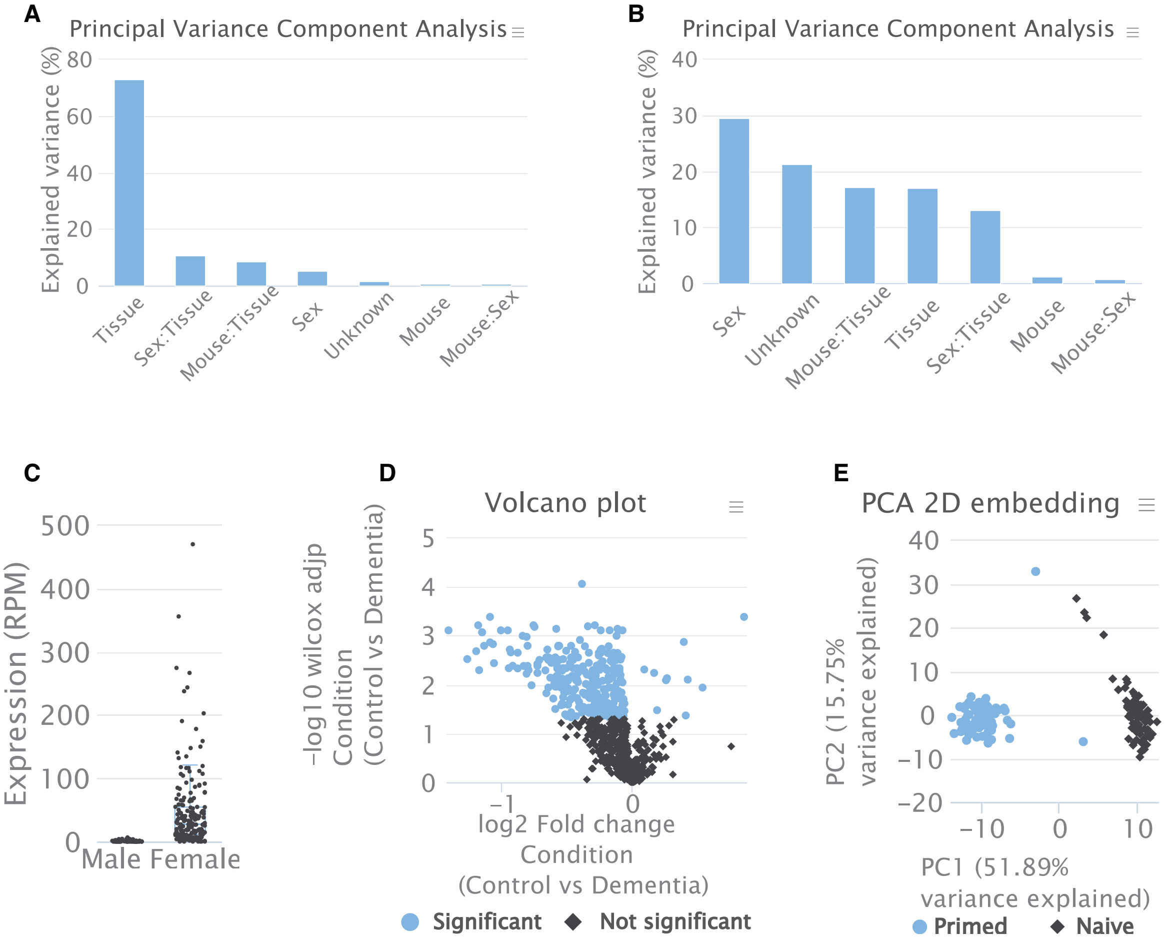


Figure 4. Downstream analysis results for the use cases. (**A**) Principal Variance Component Analysis based on the miRNA expression matrix, showing most of the variance explained by the mouse tissue. (**B**) Principal Variance Component Analysis based on the circRNA expression matrix, showing most of the variance explained by the mouse sex. (**C**) Reads per million (RPM) normalized expression of mmu_circ_0004351. (**D**) Volcano plot showing significantly deregulated miRNAs in dementia patients. (**E**) PCA embedding of the miRNA expression matrix showing a perfect separation between primed and naïve hESCs.

PCA and UMAP embeddings, as well as the sample correlation and expression clustering. The circRNA showing the most significant change between male and female mice was mmu_circ_0004351, with a fold change of 17.5 and an FDR adjusted Wilcoxon Mann-Whitney *P*-value of 4.6 × 10^{-32} (Figure 4C). miRNAs for which only 5% of the observed variance could be attributed to the mouse sex, were also partially differentially expressed, and have been previously reported in literature for single tissues or other species, such as mmu-miR-27a-3p and miR-27b-3p (44,45) and miR-16-5p and miR-21-3p (46).

Use Case (2): Differential sncRNA expression in neurodegeneration

We previously analyzed the role of microRNAs in Alzheimer's disease (4,47,48). Moreover, the examples shown above are from Alzheimer's patients and controls (Figures 2 and 3). Recently, we published a data set composed of dementia patients, including Alzheimer's disease and controls with a new technology, termed CoolMPS (49). To demonstrate the functionality of miRMaster for this sequencing assay, we analyzed the 216 case and control samples. The pre-processing page shows that all samples have been sequenced with more than 13 million reads, obtaining a median of 27.5 million for all samples. The genome mapping statistics show high mapping rates with a median of 95.5% and only few samples exhibiting a mapping rate below 90%. Since the RNAs of this data set were size selected to enrich for miRNAs, we observe the expected high mapping rates to miRNAs as well, with a median of 90.0%. The batch effect analysis suggests that most of the variation cannot be explained by any of the provided variables, followed by the age group and condition for most RNA classes. This is reflected by the sample embeddings as well as sample and expression clusterings, since no separation according to any of the provided annotations can be achieved. The differential expression results of the RNA classes show a general trend of over-expression in dementia patients, since most RNAs are down-regulated in control patients. We find that 695 miRNAs were expressed with at least 3 reads in more than 50% of either control or dementia patients and that 270 miRNAs were significantly de-regulated with an adjusted Wilcoxon Mann-Whitney *P*-value below 0.05 (Figure 4D). Upon triggering an enrichment analysis with miEAA 2.0, we find 1,790 affected categories comprising known dementia and Alzheimer's disease related pathways such as the positive regulation of endoplasmic reticulum unfolded protein response (FDR adjusted *P*-value of 0.003, (50)) and Rab GTPase binding (FDR adjusted *P*-value 0.001, (51)).

Use Case (3): analysis of data with unique molecular identifiers

Due to the increasing popularity of single cell small non-coding RNA sequencing, we also demonstrate the capabilities of miRMaster on 168 primed and naïve hESCs sequenced with a UMI-based protocol presented by Faridani *et al.* (52). Based on the pre-processing of miRMaster we find that the sequencing depth distribution is similar between primed and naïve hESCs and ranges from 1 million reads up to 44 million with a median of 4.2 million. As expected for small input protocols, and especially in the context of single cell miRNA-seq, only about 50% of the reads were kept after adapter trimming and quality filtering, which can often be attributed to adapter dimerization. The genome mapping statistics show a large variability going from 4.4% up to 88.3%, where on median 0.9% of the reads mapped against microRNAs. The most represented RNA class are rRNAs (median 6.3%) followed by snoRNAs with a median of 3.8%. The batch effect analysis shows that most of the variance in miRNA and snoRNA counts is associated with the cell state (76.3% and 82.9%), in contrast to the other RNA classes, where most of the variance cannot be explained by the cell state. This separation is clearly displayed by the miRNA PCA embedding, where the first component almost perfectly splits the primed from the naïve cells (Figure 4E). The miRNA sample correlation matrix shows high similarity (>0.7) between the cells of each state, while the correlation between the groups is in the range between 0.2 and –0.2. The correlation clustering shows similar patterns for the snoRNAs, however, the correlation coefficients inside the same cell state are higher and in-between cell states vary in the range of 0.7 and 0.85. As highlighted in the publication by Faridani *et al.*, the most up-regulated miRNA in naïve hESCs is hsa-miR-371a-5p (fold change of 797, FDR adjusted *P*-value of 1.11×10^{-27}), whereas the most down-regulated miRNA is hsa-miR-363–3p (fold change of 0.003, FDR adjusted *P*-value of 1.07×10^{-29}). It is evident that the current comparably shallow single cell small RNA data sets are not sufficient to detect all the non-coding RNA molecules present in a single cell. Especially the lower abundant non-coding RNA classes might not be sufficiently represented given the current experimental limitations. Nonetheless, miRMaster 2 identifies the molecules that are present in the sequencing data, representing a functional single cell non-coding RNA analysis pipeline.

CONCLUSION AND FUTURE DIRECTION

One of the most important novel features is certainly the support for multiple species, with which we expect to largen our userbase. Furthermore, we added circRNAs as new RNA class. Moreover, the scope of the analysis modules has been widened and we offer new data analysis aspects, such as (i) sample correlation, (ii) expression clustering, (iii) embedding, (iv) batch effect assessment and (v) differential expression analyses. Finally, we extended the adapter trimming procedure to support more major sncRNA-seq library protocols and other custom protocols. In the same context, we implemented support for UMI based analysis, thereby making the tool ready for single cell sncRNA data. In addition to the new features, we performed a major update of the underlying data bases to their current standards, improved the user-experience as well as the representation of results in tables and as interactive plots.

While we incorporated a lot of feedback from researchers that used miRMaster in this update and therefore implemented new features that were missing also from our own experience, we still see potential to further develop the usability and scope of miRMaster. One aspect that we plan to improve is automated quality control. This includes a prediction of the uploaded sample type. By using annotated data, the solid tissue or body fluid from which the data were generated can be predicted with over 90% accuracy (43). A further step is to improve the functional support for viruses and bacteria. Originally, we implemented this step for detecting contamination or reporting the presence of exogenous species. While we tested this feature only using expression data from Myxobacteria, we now aim to design a dedicated analysis module for sRNAs from microorganisms. In our ambition to develop an AI-based quality control we plan to implement automated outlier detection and propose users to perform the computational analyses after excluding flagged outliers.

To further advance the development of miRMaster, we encourage the community to continue providing us constant feedback as well as to propose new features that are of broad interest.

DATA AVAILABILITY

miRMaster 2 is freely available at https://www.ccb.uni-saarland.de/mirmaster2.

FUNDING

Michael J. Fox foundation [14446]; Schaller-Nikolich Foundation. Funding for open access charge: Michael J. Fox Foundation for Parkinson's Research.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2018) Metazoan microRNAs. *Cell*, **173**, 20–51.
2. Stavast,C.J. and Erkeland,S.J. (2019) The non-canonical aspects of microRNAs: many roads to gene regulation. *Cells*, **8**, 1465.
3. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Wurstle,M.L., Hubenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.

4. Fehlmann,T., Meese,E. and Keller,A. (2017) Exploring ncRNAs in Alzheimer's disease by miRMaster. *Oncotarget*, **8**, 3771–3772.
5. Chen,L., Heikkinen,L., Wang,C., Yang,Y., Sun,H. and Wong,G. (2019) Trends in the development of miRNA bioinformatics tools. *Brief. Bioinform.*, **20**, 1836–1852.
6. Schmartz,G.P., Kern,F., Fehlmann,T., Wagner,V., Fromm,B. and Keller,A. (2020) Encyclopedia of tools for the analysis of miRNA isoforms. *Brief. Bioinform.*, doi:10.1093/bib/bbaa346.
7. Kesharwani,R.K., Chiesa,M., Bellazzi,R. and Colombo,G.I. (2019) CBS-miRSeq: a comprehensive tool for accurate and extensive analyses of microRNA-sequencing data. *Comput. Biol. Med.*, **110**, 234–243.
8. Kanke,M., Baran-Gale,J., Villanueva,J. and Sethupathy,P. (2016) miRquant 2.0: an expanded tool for accurate annotation and quantification of MicroRNAs and their isomiRs from small RNA-sequencing Data. *J Integr Bioinform*, **13**, 307.
9. Aparicio-Puerta,E., Lebron,R., Rueda,A., Gomez-Martin,C., Giannoukakos,S., Jaspez,D., Medina,J.M., Zubkovic,A., Jurak,I., Fromm,B. *et al.* (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res.*, **47**, W530–W535.
10. Vitsios,D.M. and Enright,A.J. (2015) Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*, **31**, 3365–3367.
11. Shi,J., Dong,M., Li,L., Liu,L., Luz-Madrigal,A., Tsonis,P.A., Del Rio-Tsonis,K. and Liang,C. (2015) mirPRo-a novel standalone program for differential expression and variation analysis of miRNAs. *Sci. Rep.*, **5**, 14617.
12. Lu,Y., Baras,A.S. and Halushka,M.K. (2018) miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics*, **19**, 275.
13. Wan,C., Gao,J., Zhang,H., Jiang,X., Zang,Q., Ban,R., Zhang,Y. and Shi,Q. (2017) CPSS 2.0: a computational platform update for the analysis of small RNA sequencing data. *Bioinformatics*, **33**, 3289–3291.
14. Kuksa,P.P., Amlie-Wolf,A., Katanic,Z., Valladares,O., Wang,L.S. and Leung,Y.Y. (2018) SPAR: small RNA-seq portal for analysis of sequencing experiments. *Nucleic Acids Res.*, **46**, W36–W42.
15. Lott,S.C., Schafer,R.A., Mann,M., Backofen,R., Hess,W.R., Voss,B. and Georg,J. (2018) GLASSgo - automated and reliable detection of sRNA homologs from a single input sequence. *Front Genet*, **9**, 124.
16. Seguin,J., Otten,P., Baerlocher,L., Farinelli,L. and Pooggin,M.M. (2016) MISIS-2: a bioinformatics tool for in-depth analysis of small RNAs and representation of consensus master genome in viral quasispecies. *J. Virol. Methods*, **233**, 37–40.
17. Ebert,M.S., Neilson,J.R. and Sharp,P.A. (2007) MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat. Methods*, **4**, 721–726.
18. Zhang,J., Liu,L., Xu,T., Xie,Y., Zhao,C., Li,J. and Le,T.D. (2019) miRspongeR: an R/Bioconductor package for the identification and analysis of miRNA sponge interaction networks and modules. *BMC Bioinformatics*, **20**, 235.
19. Desvignes,T., Batzel,P., Sydes,J., Eames,B.F. and Postlethwait,J.H. (2019) miRNA analysis with Prost! reveals evolutionary conservation of organ-enriched expression and post-transcriptional modifications in three-spined stickleback and zebrafish. *Sci. Rep.*, **9**, 3913.
20. Zhang,Y., Zang,Q., Zhang,H., Ban,R., Yang,Y., Iqbal,F., Li,A. and Shi,Q. (2016) DeAnnIso: a tool for online detection and annotation of isomiRs from small RNA sequencing data. *Nucleic Acids Res.*, **44**, W166–W175.
21. Chen,Y., Ruan,Z.R., Wang,Y., Huang,Q., Xue,M.Q., Zhou,X.L. and Wang,E.D. (2018) A threonyl-tRNA synthetase-like protein has tRNA aminoacylation and editing activities. *Nucleic Acids Res.*, **46**, 3643–3656.
22. Winek,K., Lobentanzer,S., Nadorp,B., Dubnov,S., Dames,C., Jagdmann,S., Moshitzky,G., Hotter,B., Meisel,C., Greenberg,D.S. *et al.* (2020) Transfer RNA fragments replace microRNA regulators of the cholinergic poststroke immune blockade. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 32606–32616.
23. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
24. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
25. Consortium,R.N. (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.*, **49**, D212–D220.
26. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
27. Glazar,P., Papavasileiou,P. and Rajewsky,N. (2014) circBase: a database for circular RNAs. *RNA*, **20**, 1666–1670.
28. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
29. Koster,J. and Rahmann,S. (2018) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **34**, 3600.
30. Fernandez,N.F., Gundersen,G.W., Rahman,A., Grimes,M.L., Rikova,K., Hornbeck,P. and Ma'ayan,A. (2017) Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Sci Data*, **4**, 170151.
31. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
32. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
33. Desvignes,T., Loher,P., Eilbeck,K., Ma,J., Urgese,G., Fromm,B., Sydes,J., Aparicio-Puerta,E., Barrera,V., Espin,R. *et al.* (2020) Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API. *Bioinformatics*, **36**, 698–703.
34. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
35. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grasser,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
36. Solomon,J., Kern,F., Fehlmann,T., Meese,E. and Keller,A. (2020) HumiR: web services, tools and databases for exploring human microRNA Data. *Biomolecules*, **10**, 1576.
37. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
38. Kern,F., Fehlmann,T., Solomon,J., Schwed,L., Grammes,N., Backes,C., Van Keuren-Jensen,K., Craig,D.W., Meese,E. and Keller,A. (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res.*, **48**, W521–W528.
39. Kern,F., Amand,J., Senatorov,I., Isakova,A., Backes,C., Meese,E., Keller,A. and Fehlmann,T. (2020) miRSwitch: detecting microRNA arm shift and switch events. *Nucleic Acids Res.*, **48**, W268–W274.
40. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) MiRTargetLink–miRNAs, genes and interaction networks. *Int. J. Mol. Sci.*, **17**, 564.
41. Backes,C., Sedaghat-Hamedani,F., Frese,K., Hart,M., Ludwig,N., Meder,B., Meese,E. and Keller,A. (2016) Bias in high-throughput analysis of miRNAs and implications for biomarker studies. *Anal. Chem.*, **88**, 2088–2095.
42. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grasser,F.A., Lenhof,H.P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, **47**, 3353–3364.
43. Isakova,A., Fehlmann,T., Keller,A. and Quake,S.R. (2020) A mouse tissue atlas of small noncoding RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 25634–25645.
44. Guo,D., Ye,Y., Qi,J., Tan,X., Zhang,Y., Ma,Y. and Li,Y. (2017) Age and sex differences in microRNAs expression during the process of thymus aging. *Acta Biochim. Biophys. Sin. (Shanghai)*, **49**, 409–419.
45. Hermenean,A., Trotta,M.C., Gharbia,S., Hermenean,A.G., Peteu,V.E., Balta,C., Cotoraci,C., Gesualdo,C., Rossi,S., Gherghiceanu,M. *et al.* (2020) Changes in retinal structure and ultrastructure in the aged mice correlate with differences in the expression of selected retinal miRNAs. *Front Pharmacol*, **11**, 593514.

46. Hasakova,K., Bezakova,J., Vician,M., Reis,R., Zeman,M. and Herichova,I. (2017) Gender-dependent expression of leading and passenger strand of miR-21 and miR-16 in human colorectal cancer and adjacent colonic tissues. *Physiol. Res.*, **66**, S575–S582.
47. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement*, **12**, 565–576.
48. Ludwig,N., Fehlmann,T., Kern,F., Gogol,M., Maetzler,W., Deutscher,S., Gurlit,S., Schulte,C., von Thaler,A.K., Deuschle,C. *et al.* (2019) Machine learning to detect Alzheimer's disease from circulating non-coding RNAs. *Genomics Proteomics Bioinformatics*, **17**, 430–440.
49. Li,Y., Fehlmann,T., Borcherding,A., Drmanac,S., Liu,S., Groeger,L., Xu,C., Callow,M., Villarosa,C., Jorjorian,A. *et al.* (2021) CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing. *Nucleic Acids Res.*, **49**, e10.
50. Scheper,W. and Hoozemans,J.J. (2015) The unfolded protein response in neurodegenerative diseases: a neuropathological perspective. *Acta Neuropathol.*, **130**, 315–331.
51. Zhang,X., Huang,T.Y., Yancey,J., Luo,H. and Zhang,Y.W. (2019) Role of Rab GTPases in Alzheimer's sisease. *ACS Chem. Neurosci.*, **10**, 828–838.
52. Faridani,O.R., Abdullayev,I., Hagemann-Jensen,M., Schell,J.P., Lanner,F. and Sandberg,R. (2016) Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.*, **34**, 1264–1266.

3.3 *miRSwitch: detecting microRNA arm shift and switch events*

W268–W274 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 1 May 2020*
doi: 10.1093/nar/gkaa323

miRSwitch: detecting microRNA arm shift and switch events

Fabian Kern[1], Jeremy Amand[1], Ilya Senatorov[1], Alina Isakova[2], Christina Backes[1], Eckart Meese[3], Andreas Keller[1,4,5,*] and Tobias Fehlmann[1]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA, [3]Department of Human Genetics, Saarland University, 66421 Homburg, Germany, [4]School of Medicine Office, Stanford University, Stanford, CA 94305, USA and [5]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94304, USA

Received March 25, 2020; Revised April 15, 2020; Editorial Decision April 20, 2020; Accepted April 22, 2020

ABSTRACT

Arm selection, the preferential expression of a 3′ or 5′ mature microRNA (miRNA), is a highly dynamic and tissue-specific process. Time-dependent expression shifts or switches between the arms are also relevant for human diseases. We present miRSwitch, a web server to facilitate the analysis and interpretation of arm selection events. Our species-independent tool evaluates pre-processed small non-coding RNA sequencing (sncRNA-seq) data, i.e. expression matrices or output files from miRNA quantification tools (miRDeep2, miRMaster, sRNAbench). miRSwitch highlights potential changes in the distribution of mature miRNAs from the same precursor. Group comparisons from one or several user-provided annotations (e.g. disease states) are possible. Results can be dynamically adjusted by choosing from a continuous range of highly specific to very sensitive parameters. Users can compare potential arm shifts in the provided data to a human reference map of pre-computed arm shift frequencies. We created this map from 46 tissues and 30 521 samples. As case studies we present novel arm shift information in a Alzheimer's disease biomarker data set and from a comparison of tissues in *Homo sapiens* and *Mus musculus*. In summary, miRSwitch offers a broad range of customized arm switch analyses along with comprehensive visualizations, and is freely available at: https://www.ccb.uni-saarland.de/mirswitch/.

INTRODUCTION

The non-coding parts of mammalian genomes play a major role in shaping the gene regulatory landscape (1,2). Still, these mechanisms are only understood to a limited extent. Among the many different classes of small or long non-coding RNA elements (3), which modulate the expression of genes on a transcriptional or translational level, microRNAs (miRNAs) seem to play a key role (4,5). In the biogenesis of miRNAs, from nascent pri-miRNAs to mature forms, the precursor hairpin molecules are processed by the enzymes DROSHA and DICER (6). The product of the second cleavage is an approximately 22-nucleotide long RNA duplex structure. Frequently, one arm of the RNA duplex is preferentially accumulated while the other is predominantly degraded (7). For most hairpins the dominant mature miRNA is assumed to be the functional product. Gene regulation is carried out by the association of the major form with AGO proteins for RNA-induced silencing complex (RISC) formation and successive binding to reverse complementary target sites in mRNAs, mostly within 3'-untranslated regions (8).

High-throughput data from different tissues (9), aging time-points, developmental stages, and physiological conditions (10,11) indicate, a minor proportion of mature miRNAs from the opposite arm to originate nonetheless. Formerly, these sequences have been denoted as the miR* sequence. Thermodynamic and structural properties have been postulated as drivers for the selection of the dominant arm from the processed duplex (12). Previous work also demonstrates that arm shifts are specific for tissue types and the distribution of the dominant mature and the miR* sequence can change (13,14). For several precursor hairpins, significant quantities of mature miRNAs from both arms are known (15) and both of them are biologically functional. The dominant arm represses translation by means of AGO1 and the miR* sequence by means of AGO2 (16). Currently, miRNAs are not annotated as miR* and dominant mature form anymore but precisely denoted as the −3p and −5p mature miRNA. Beyond these observations, earlier studies describe the process of selective arm switches in multiple pathological conditions (e.g. (17–22)). In 2011, Griffith-Jones *et al.* found that arm usage is encoded in the

[*]To whom correspondence should be addressed: Tel: +49 681 302 68611; Email: andreas.keller@ccb.uni-saarland.de

primary miRNA sequence, but outside the mature miRNA duplex, by analysing the miR-100/10 family in different species (16). The group also provided evidence for functional shifts in insect miRNA evolution (23). Arm switches have also been correlated to human pathologies such as breast cancer and other severe disorders (18).

Still, a systematic analysis of arm shift or switch events from high-throughput data has not been proposed yet. Here, we introduce miRSwitch, a tool to find differential arm expression from pre-processed high-throughput expression data. In the context of miRSwitch an arm *shift* is a significant enrichment of the 3′ or 5′ arm in one compared to another condition, while still preserving the identity of the dominant form. An arm *switch* denotes a more extreme scenario where in one condition either arm is the dominant and vice versa in the other condition. The supported type of input of miRSwitch ranges from expression matrices over result files from common high-throughput tools such as miRDeep2 (24), miRMaster (25) or sRNAbench (26) to prominent data sources like The Cancer Genome Atlas (TCGA). Users can upload annotation files or manually annotate the samples before running an analysis. To provide even further insights into arm shift events we facilitate a comparison to a background (reference) map of curated arm switch events in *Homo sapiens*. We generated this map from 38 252 human sncRNA-seq data sets comprising 556 billion (5.556×10^{11}) reads from 46 different tissues. To demonstrate the functionality of miRSwitch we evaluate an Alzheimer's disease data set. As second case study, we compare the frequency of arm switches in human and mouse tissues to test the hypothesis whether arm switches might be conserved across species.

MATERIALS AND METHODS

Differential arm expression detection

In a first step, miRNA-precursor pairs of the uploaded expression files are converted to their miRBase v22.1 (27) identifier using the miRBaseConverter R package (28) (v1.10.0). Next, the precursors are filtered, such that only precursors that have a miRBase identifier and two annotated miRNA arms remain. Then, we remove precursors with no miRNA exceeding a user specified threshold for the minimal number of reads. We compute the $5' - 3'$ ratio difference for every precursor in every sample. Statistical significance between two levels of one annotation variable is calculated with the Wilcoxon-rank sum and for three or more levels using a Kruskal-Wallis test. In all cases, *P*-values are corrected for multiple hypothesis testing using the Benjamini–Hochberg correction for controlling the false discovery rate. For annotation variables with two levels we also compute the area under the receiver operating characteristic curve (AUC) as effect size. Given a precursor with a 5′ miRNA expression value of e_5 and a 3′ miRNA expression value of e_3, we define $e = max(e_5, e_3)$ and $a = \frac{e}{e_5+e_3}$ as arm ratio. Also, let R be the minimum arm ratio threshold and X be the minimum miRNA reads threshold. Precursors are classified as 5′ dominant in one sample, if $a \geq R$ and $e \geq X$ and $e_5 > e_3$, 3′ dominant if $a \geq R$ and $e \geq X$ and $e_5 < e_3$, not dominant if $a < R$ and $e \geq X$, and not expressed if $e < X$. Arm switch candidates can be queried by defining an additional threshold S, requiring 5′ and 3′ dominant precursors in at least S samples.

Web server implementation

We implemented miRSwitch using a dockerized Django Web Framework (v2.2) with a MonetDB database backend (v11.35.19). As job scheduler we used the celery software (v4.3.0). To build a user frontend we used Webpack (v4.41.2) in combination with React JS (v16.12.0), Dev Extreme React Grid (v2.3.2), fornac (v1.1.0), Plotly (v1.51.3), and Highcharts (v7.2.1). The specificity and sensitivity trade-off for potential arm shift/switch candidates can be controlled by adjusting three parameters. These are the minimum arm shift ratio R, the minimal number of samples S where the threshold R needs to be exceeded and the minimal number of reads X of one miRNA arm required to compute the 5′ to 3′ ratio.

Human reference map of differential arm expression

To generate a reference map of human arm shift / switch events we collected 38 252 human sncRNA-seq samples. The data set was compiled from three different sources, the Sequence Read Archive (SRA) (29) (16,415), TCGA (30) (10 999), and samples that were made accessible by anonymous users of miRMaster (25) and who provided consent for aggregated secondary usage (10 838). Subsequently, we removed duplicated data sets that occurred after pooling the SRA and miRMaster samples. All samples were processed as previously described (31). Briefly, samples were mapped against the human genome (hg38) and discarded from further analyses in case less than 50% of reads could be aligned with Bowtie (v1.1.2) (32) while allowing no mismatches. We also discarded samples for which either at least 1% of reads mapped to coding regions or fewer than 1 million total reads were detected. After applying all filtering steps, 30 521 samples remained for consideration. For the samples from SRA, TCGA, and a subset of the miRMaster samples for which annotations were known, the annotation metadata was included in the web server. Other samples for which no tissue annotation was available were labelled as 'Unknown'. The final map is consistent with the most recent release of miRBase v22.1 and contains 961 human precursors, which are all annotated with two mature forms.

Case studies

To demonstrate the functionality of miRSwitch we performed two case studies; a human liquid biopsy biomarker study and a consideration of arm switches in human compared to mouse tissues. Previously, we obtained sncRNA-seq data from blood of Alzheimer's disease patients and controls, and validated the data using RT-qPCR (33–35). For 70 samples, reads were mapped and miRNA counts for miRBase v21 entries were quantified using miRDeep2. This case study is linked as example data set on the miRSwitch analysis page. As second case study we investigate arm switch events of the same tissues between mouse and human (doi:10.1101/430561).

RESULTS

In its essence, miRSwitch allows to search for differential miRNA arm expression between any kind of biological condition. In the following, the basic workflow as well as details and examples on the individual steps are described. The results are concluded with two case studies to demonstrate the features on real-world scenarios where new biological insights can be inferred.

Workflow of miRSwitch

The workflow of a miRSwitch analysis entails four steps, (1) the input specification; (2) the main analysis functionality; (3) an optional comparison to the human reference map and (4) the representation and export of results (Figure 1). Each of the steps is described below in more detail. In-depth guidance using illustrated online tutorials as well as video tutorials is accessible from the miRSwitch website.

Step 1: Input specification. As the first step, miRSwitch provides a very flexible user interface to transfer data to the server. For example, the user provides miRNA expression data and an annotation matrix file (.csv, .tsv). The expression data can be uploaded either as a expression matrix file or the plain output from common miRNA discovery and quantification tools applied up-stream. The supported pre-processing tools include miRDeep2, miRMaster and sRNAbench. Unmodified data files from TCGA can be uploaded as well. Next, the interface automatically extracts metadata from the uploaded file(s) to be displayed in an interactive sample table. Finally, the user can check and modify any derived annotations. In any case, annotation files are optional and annotation variables can be entered manually instead. Most importantly, the first column is expected to match the sample IDs exactly. In addition to an example study that can be executed from the analysis tab, a template expression matrix and miRDeep2 project files are provided in the tutorial section.

Step 2: Arm expression analysis. From an expression matrix, 5′ and 3′ ratios are computed for each miRNA precursor with two mature forms mapped. Also, the main parameters defining the strength and frequency of arm shift/switch events can be modified. These include the threshold ratio of the 5′ or 3′ arm, the number of reads that have to match to at least one corresponding mature form for each observation and the number of samples that show a respective ratio. Interactive charts and a table of precursors and their classification with respect to the arm dominance are shown on the general analysis page.

Given sample annotations, miRSwitch also performs group comparisons. The user can access the respective results from the 'Annotation' tab (for convenience reasons the tab is always named according to the information provided by the user) on the main results page. This tab presents significance values and AUCs for the group comparisons, as well as embeddings of the samples from dimension reduction methods, highlighted according to their group. miRSwitch then computes graphical representations and spreadsheets for all results (see module 4). Since the parameter choice is overall crucial and researchers may want to use their own specific or sensitive parameters to determine whether a miRNA is able to perform an arm switch, we enable real-time parameter selection and filtering of the respective miRNAs in the web interface.

Step 3: Reference map. To provide further insights into arm shifts in *H. sapiens* we integrated a reference map of arm shift events in different solid tissues and bio fluids. We collected 38,252 sncRNA data sets from three different sources (SRA (29), TCGA (30) and data collected by our tool miRMaster (25)). These data sets contain a total of 556 billion (5.556×10^{11}) sequencing reads and can be annotated for 46 different tissue types. After a stringent quality filtering to exclude data with almost no reads mapping to *Homo sapiens* or containing other sequences but not sncRNAs, a total of 30 521 remained.

Three scenarios from the human reference map demonstrate the importance of the main analysis parameters (cf. Materials and Methods & Results, Step 2). First, to obtain a very specific view, we set the arm ratio of 3′ or 5′ to be at least 80%, at least one miRNA needs to express 1000 reads and at least 200 experiments have to show a dominant 5′ expression and another 200 a dominant 3′ expression for the mature miRNAs. For this parameter set (80, 1000, 200), 52 precursors that perform an arm switch are identified, with the most variable being hsa-mir-193a, hsa-mir-30e, hsa-let-7d, hsa-mir-144, hsa-mir-361 and hsa-mir-423. If we alter the parameter set to be less specific (70, 500, 100), already 108 precursors with potential arm switches are identified. Finally, we test a very sensitive parameter set (60, 200, 20). Here, miRSwitch reports 256 precursors with potential arm switches.

Step 4: Results representation and export. The fourth module is the representation of results generated by the analysis step (2) or extracted from the reference map in step (3). Generally, three different types of results pages are available. Two of them for user provided input and one for the background map. Additionally, detailed results for individual precursors can be viewed in both modules. If an own data set is evaluated the first tab covers general aspects. Aggregated information on how many samples and which annotations were processed, how many precursors were expressed, and how many reads per sample were available. Bubble plots represent how many miRNAs per annotation on the two arms were found and the arm distribution across samples is shown as heat map. Following this general information the user can adjust the parameters for defining arm shift events. The following bar graphs and tables are adjusted dynamically. Here, users can filter or sort the precursors. From the table, single precursors can be selected and detailed information for these candidates are provided. These include links to external databases (miRBase (27) and miRCarta (15)), the sequence and structure, a pie chart on the arm distribution and detailed distribution per annotation group. All information can be displayed either as percentages or absolute values in the bar diagrams. Finally, the information in which samples the miRNA precursor was expressed on both arms is available as interactive table. Next, the annotation tab contains more details about the comparisons of annotation variables. First, sample em-

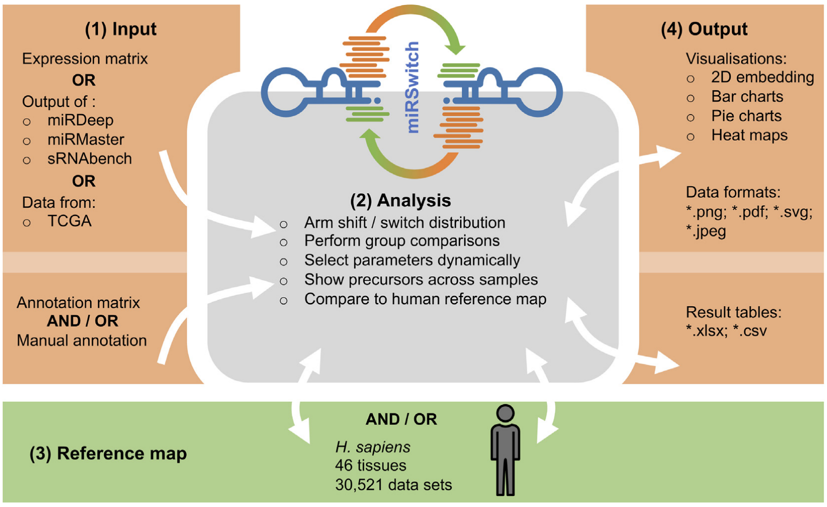


Figure 1. The miRSwitch workflow. miRSwitch consists of four steps (modules), (1) input, (2) miRSwitch analysis core, (3) reference map and (4) output. The arrows denote possible interactions between those. From the input module data are transferred unidirectional to the core module. Between the core and the background map as well as results module bidirectional communication allows to adjust the results dynamically if parameters to define arm switches are changed by the user. All obtained results are easily exportable in common data table formats and plot graphics.

beddings from Uniform Manifold Approximation and Projection (UMAP) (arXiv:1802.03426) and Principal Components Analysis (PCA) (36) of the 5′ − 3′ ratio matrix are provided. Here, color and shape of the points represent the annotation levels for the respective samples. Then, *P*-values for the difference of the 3′ and 5′ ratio are computed and provided as raw- and adjusted significance values. Another output in the interactive result table that allows sorting and filtering is the AUC value. For each miRNA, the 3′ and 5′ distribution per annotation group is available as dodged violin plot.

Finally, the *human reference* tab contains all data collected for the background map. The representation is similar to the general results obtained after starting a custom arm ratio analysis. This facilitates an easy comparison. For each precursor, the 5′ and 3′ miRNA expression overall and per tissue are shown. The detailed information for each miRNA precursor is then identical to the results provided by the analysis of own data. Moreover, distribution plots, pie charts and bar charts are computed to get insights into the distribution of the two mature arms of detected precursors. Figure 2 presents a typical example for a 5′ dominant miRNA precursor (mir-142) and a 3′ dominant miRNA precursor (mir-144).

miRSwitch aims to provide broad export flexibility for further down-stream usage. This includes the support of the most relevant image formats covering vector graphics and raster graphics (jpeg, svg, pdf and png). Furthermore detailed result files are also available as spreadsheets. This covers the excel data format as well as the tsv plain text file format.

Case study 1: Alzheimer's disease

In the first case study we consider previously published sncRNA-seq data from whole-blood samples of Alzheimer's disease patients and controls. The case study is also provided as example analysis on the miRSwitch homepage. A total of 70 sample files from miRDeep2 are processed in real-time when loading the example and results are available after ∼20 s. As first result the web server reports 2784 miRNAs and 1855 precursors in miRBase v22.1 of which 748 are expressed with at least one read and have two known mature forms. From the annotation metadata the 'Alzheimer' and 'Control' labels were identified. Using the default parameter set, our tool reports 10 candidates with arm shift and 3′ dominance as well as three candidates with 5′ dominance. Further, the results view indicates that in Alzheimer's disease fewer 5′ and more 3′ mature miRNAs are expressed as compared to controls (Figure 3A). The heat map points to one potential outlier sample (SRR837506, data not shown). With the parameter set (80, 5, 3), miRSwitch identifies 7 precursors with 3′ dominant arm (mir-340, mir-199b, mir-29c, mir-6859-1, mir-6859-2, mir-6859-3 and mir-6859-4) and one with a 5′ dominant arm (mir-548h-4). For mir-340, 9% of the samples don't show an expression, 49% do not show a dominant arm, 30% are 3′ dominant and 13% are 5′ dominant (Figure 3B). Dividing this into the two groups of patients and controls we find that except for one sample all samples with 3′ dominant mir-340 come from Alzheimer's disease patients while control samples are enriched for 5′ dominant mir-340 expression (Figure 3C). The distribution plot that shows the difference in ratios for controls (left) and patients (right) confirms this on a per sample basis (Figure 3D). Ultimately, the dodged violin distribution plot demonstrates that in case of mir-340 not only an arm shift but an arm switch between the two groups of samples can be observed (Figure 3E).

Case study 2: Arm shift events in *M. musculus* and *H. sapiens*

As second case study we analysed miRNAs from a mouse sncRNA tissue atlas (doi:10.1101/430561) to assess the potential (dis-)similarity of arm shifts in mice and human. To

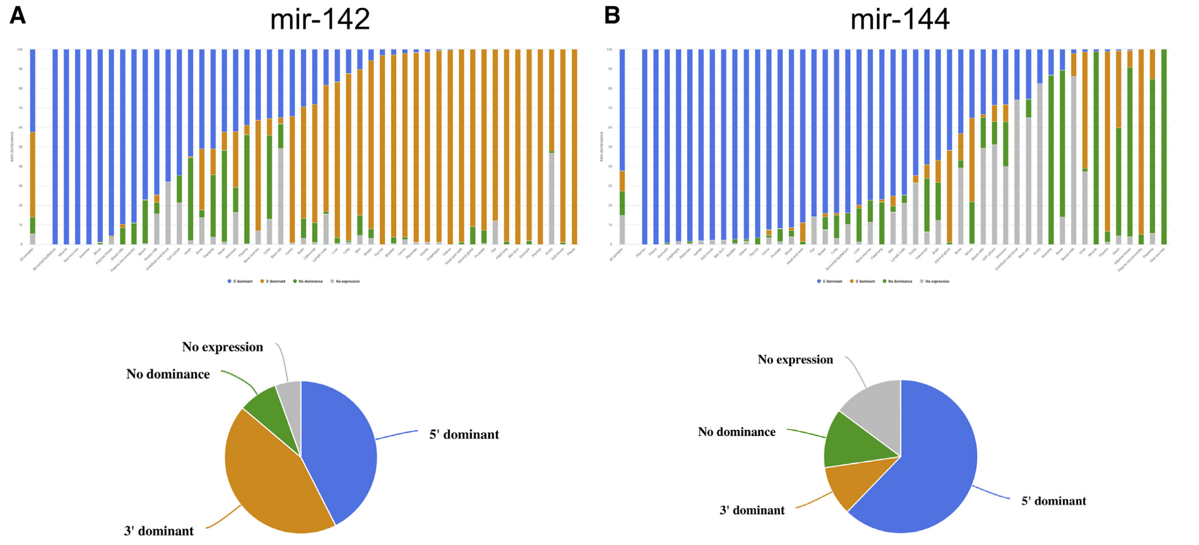


Figure 2. Example results from the human samples in the background map of arm shift and arm switch events. (**A**) For hsa-mir-142 the distribution of 3′ dominant, 5′ dominant, no dominance, and not expressed are presented for all samples as bar and pie chart. The same classification is also presented for each tissue. For ileal mucosa or blood, the 5′ form is clearly dominant, for thymus or pleura, the 3′ form dominates. (**B**) For hsa-mir-144 the 5′ mature form is clearly more abundant as compared to the 3′ form but in several tissues such as the muscle, no form is dominant.

this end, we showcase the value of the human reference map feature to perform the species comparison. We restricted the focus to solid organs included in both data sets. For example, mir-141 is 3′ dominant in both species, but a higher expression of the 5′ arm was observed for both organisms in testis. Also, mir-26b was largely 5′ dominant in *M. musculus*, only the bone marrow showed expression of both arms. Interestingly, also the human data showed this pattern, although with lower 3′ expression ratios. For mir-106b both organism indicated expression of both arms. In this case, however, a dominant 5′ arm in the human heart was not discovered in mouse samples. mir-337 was mostly 3′ dominant. Brain samples of both organisms however indicate an increased 5′ expression. Although the direct comparison of the mouse tissue data set and the human reference map is biased in its nature, since the latter contains three orders of magnitude more samples from different conditions, we found evidence for many human miRNA arm selection events also in the mice.

DISCUSSION

Arm shifts and arm switch events have a high impact in many research scenarios, while possible down-stream effects are still underestimated. For example, previous results demonstrate an altered arm distribution between affected and unaffected individuals. Such events have been observed, e.g. for breast cancer (18), gastric cancer (22) or prostate cancer (37). Also the cause of the differences in arm distribution between cell types, developmental stages, and in diseases have been explored only to a limited extent (38,39). One likely reason that arm switches have been widely neglected so far, is the missing functionality for arm switch tailored analyses in many standard tools, including our own sncRNA-seq analysis tool miRMaster. The primary goal of this work was to make comprehensive arm switch analyses for any kind of experiment like microarrays, high-throughput sequencing, and RT-qPCR data available to a broader research community. Additionally, well interpretable output is delivered back to the user as interactive graphics and tables. Moreover, miRSwitch was designed to scale well with an increasing number of samples. The example provided online (70 samples) is analysed in approximately 20 seconds. We further tested our tool on one of the currently largest sequencing studies, profiling of almost 4,400 miRNA samples from the Parkinson's Progression Markers Initiative (PPMI) for which the job results were obtained in ∼10 min. Still, one limiting step might be the data upload that heavily depends on the connection quality. Nevertheless, the matrix format is very efficient and can be recommended in case the internet connection transfer rate is a bottleneck.

A challenging task of the custom arm switch analysis functionality is to calculate a measure of accuracy of predicted events. In fact, the accuracy depends on several aspects, for example the quantity and quality of samples provided by the user, or whether candidate arm shifts in the considered scenarios occur frequently or only exist barely. Typically encountered for high-throughput studies, this calls for dedicated validation experiments, either with a second independent high-throughput data set or with a specific low-throughput validation assay. Since this would impose an effort for potential users not to be neglected, we implemented a reference map containing arm switch events for human, as it is currently in the focus of research on differential miRNA arm expression. Here, users can check whether their results have already been discovered in any of the previously screened tissues. We consider the map as on-going effort and which will be developed towards a more complete database of arm shift events. Future extensions should cover

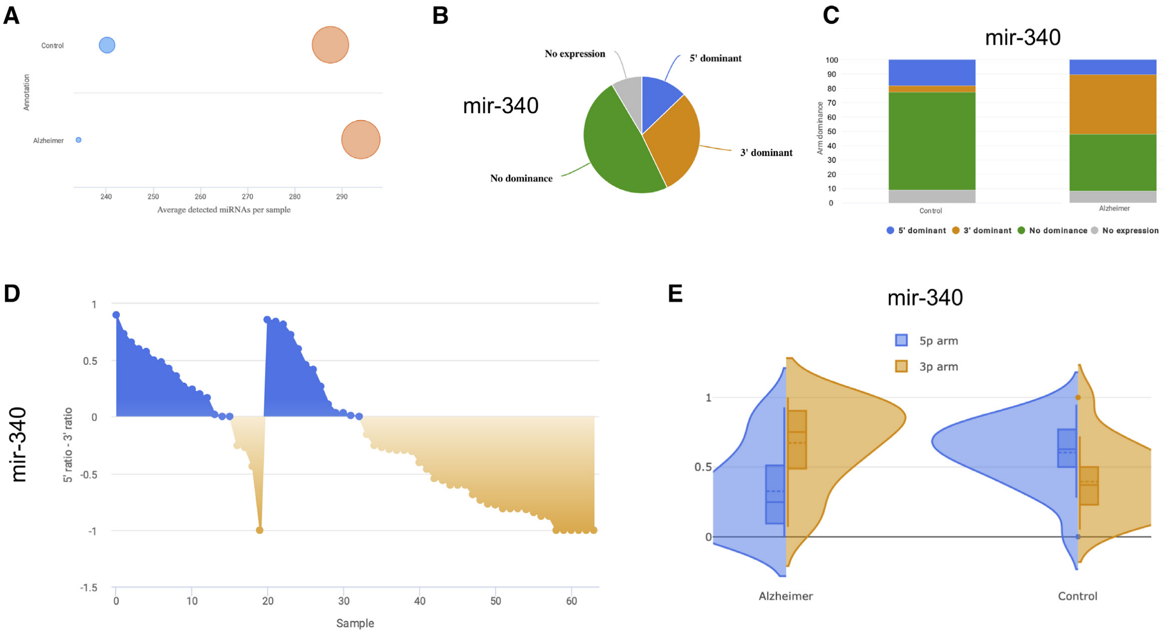


Figure 3. Results of the Alzheimer's disease case study. (**A**) Bubble plot showing the number of detected mature miRNAs from both arms and for each annotation. (**B**) Representative pie chart for mir-340 showing in how many samples it is not expressed, shows no arm preference, or is 3′ or 5′ dominant. Plots on the results page are interactive and by hovering over details are displayed. (**C**) Bar chart that splits the information from the pie chart in panel (B) into the annotation levels. (**D**) Distribution chart presenting details on the 5′ − 3′ difference for the provided groups of samples, here controls (left) and Alzheimer's disease patients (right). For the latter, a strong enrichment of the 3′ mature form is visible. (**E**) Back-to-back distribution of the arm expression in the provided annotation groups. mir-340 is an arm switch miRNA in Alzheimer's disease. In the disease it displays higher expression of the 3′ arm while in controls the 5′ arm is more abundant.

three main aspects: First, we will include more samples for *H. sapiens* to cover more tissues and add data from other sequencing technologies like cPAS Sequencing By Synthesis (40). Second, we will enlarge the set of miRNAs with two annotated forms using novel miRNA candidates. As a third extension we plan to support other organisms, e.g. mouse or rat.

Proper annotation of miRNAs is still a challenging issue. However, not all mature miRNAs might be available in public databases. Several studies pointed out the erroneous nature of many mature miRNAs in miRBase (15,31,41,42). The current human reference map is consistent with miRBase v22.1. We aim to augment this setup by offering three types of miRNA annotations. First, a very specific set could be high-confidence annotations from miRBase. The second one could comprise all annotations from miRBase and the third novel miRNA candidates from other databases, e.g. miRCarta. For the custom analysis functionality, this extension could facilitate the discovery of *de novo* arm switching events, e.g. in a disease context as demonstrated previously for gastric cancer (43). Another challenge is the underlying technology. It is well known that detection of miRNAs, similar to other non coding RNAs, mRNAs or proteins, varies depending on the experimental techniques and protocols (15,40–42). Most of the data sets used in the reference map stem from Illumina Sequencing By Synthesis instruments, which may show e.g. a ligation bias (44). As a consequence, not all potential arm shift events will be discovered due to respective bias. Secondly, the comparison between the user data set and the uploaded data might be compromised and detected differences might occur due to the difference in technologies. To assess this issue, we plan to grow the reference map further in the directions outlined above, to promote a less biased comparison and evaluation procedure.

In conclusion, we herein present a very comprehensive web server that facilitates the evaluation of arm shift and arm switch events across different species.

DATA AVAILABILITY

miRSwitch is freely available at https://www.ccb.uni-saarland.de/mirswitch. No login is required. The data for the case studies are available from the Gene Expression Omnibus (GEO) under accession numbers GSE46579 and GSE119661.

ACKNOWLEDGEMENTS

We acknowledge the work of the Meese Lab, generating the data set on Alzheimer's disease.

FUNDING

Michael J. Fox foundation [14446]; the development of the miRSwitch resource has been funded by Saarland University and the state government. Funding for open access charge: Michael J. Fox Foundation for Parkinson's Research.

Conflict of interest statement. None declared.

REFERENCES

1. Morris,K.V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.*, **15**, 423–437.
2. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet*, **15**, R17–R29.
3. Costa,F.F. (2010) Non-coding RNAs: Meet thy masters. *BioEssays*, **32**, 599–608.
4. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
5. He,L. and Hannon,G.J. (2004) MicroRNAs: Small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.
6. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
7. Hutvagner,G. (2005) Small RNA asymmetry in RNAi: Function in RISC assembly and gene regulation. *FEBS Lett.*, **579**, 5850–5857.
8. Valencia-Sanchez,M.A., Liu,J., Hannon,G.J. and Parker,R. (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Gene. Dev.*, **20**, 515–524.
9. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stähler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
10. Keller,A., Leidinger,P., Bauer,A., Elsharawy,A., Haas,J., Backes,C., Wendschlag,A., Giese,N., Tjaden,C., Ott,K. *et al.* (2011) Toward the blood-borne miRNome of human diseases. *Nat. Methods*, **8**, 841–843.
11. Hecksteden,A., Leidinger,P., Backes,C., Rheinheimer,S., Pfeiffer,M., Ferrauti,A., Kellmann,M., Sedaghat,F., Meder,B., Meese,E. *et al.* (2016) miRNAs and sports: Tracking training status and potentially confounding diagnoses. *J. Transl. Med.*, **14**, 219.
12. Meijer,H.A., Smith,E.M. and Bushell,M. (2014) Regulation of miRNA strand selection: follow the leader? *Biochem. Soc. Trans.*, **42**, 1135–1140.
13. Guo,L. and Lu,Z. (2010) Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data. *Comput. Biol. Chem*, **34**, 165–171.
14. Kuo,W.T., Su,M.W., Lee,Y.L., Chen,C.H., Wu,C.W., Fang,W.L., Huang,K.H. and Lin,W.C. (2015) Bioinformatic Interrogation of 5p-arm and 3p-arm Specific miRNA Expression Using TCGA Datasets. *J. Clin. Med.*, **4**, 1798–1814.
15. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.P., Meese,E. and Keller,A. (2018) MiRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
16. Griffiths-Jones,S., Hui,J.H., Marco,A. and Ronshaugen,M. (2011) MicroRNA evolution by arm switching. *EMBO Rep.*, **12**, 172–177.
17. Lin,M.H., Chen,Y.Z., Lee,M.Y., Weng,K.P., Chang,H.T., Yu,S.Y., Dong,B.J., Kuo,F.R., Hung,L.T., Liu,L.F. *et al.* (2018) Comprehensive identification of microRNA arm selection preference in lung cancer: MiR-324-5p and -3p serve oncogenic functions in lung cancer. *Oncol. Lett.*, **15**, 9818–9826.
18. Tsai,K.W., Leung,C.M., Lo,Y.H., Chen,T.W., Chan,W.C., Yu,S.Y., Tu,Y.T., Lam,H.C., Li,S.C., Ger,L.P. *et al.* (2016) Arm selection preference of MicroRNA-193a varies in breast cancer. *Sci. Rep.-UK*, **6**, 28176.
19. Guo,L., Yu,J., Yu,H., Zhao,Y., Chen,S., Xu,C. and Chen,F. (2015) Evolutionary and expression analysis of miR-#-5p and miR-#-3p at the miRNAs/isomiRs levels. *Biomed. Res. Int.*, **2015**, 168358.
20. Hu,W., Wang,T., Yue,E., Zheng,S. and Xu,J.H. (2014) Flexible microRNA arm selection in rice. *Biochem. Biophys. Res. Commun.*, **447**, 526–530.
21. Guo,L., Zhang,H., Zhao,Y., Yang,S. and Chen,F. (2014) Selected isomiR expression profiles via arm switching? *Gene*, **533**, 149–155.
22. Li,S.C., Liao,Y.L., Ho,M.R., Tsai,K.W., Lai,C.H. and Lin,W.C. (2012) MiRNA arm selection and isomiR distribution in gastric cancer. *Ser. Adv. Bioinformatics Computat. Biol.*, **13**, S13.
23. Marco,A., Hui,J.H., Ronshaugen,M. and Griffiths-Jones,S. (2010) Functional shifts in insect microRNA evolution. *Genome Biol. Evol.*, **2**, 686–696.
24. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
25. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Wurstle,M.L., Hubenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.
26. Aparicio-Puerta,E., Lebrón,R., Rueda,A., Gómez-Martín,C., Giannoukakos,S., Jaspez,D., Medina,J.M., Zubkovic,A., Jurak,I., Fromm,B. *et al.* (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res.*, **47**, W530–W535.
27. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) MiRBase: From microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
28. Xu,T., Su,N., Liu,L., Zhang,J., Wang,H., Zhang,W., Gui,J., Yu,K., Li,J. and Le,T.D. (2018) miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. *BMC Bioinformatics*, **19**, 514.
29. Leinonen,R., Sugawara,H., Shumway,M and International Nucleotide Sequence Database Collaboration (2010) The Sequence Read Archive. *Nucleic Acids Res.*, **39**, D19–D21.
30. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Sander,C., Stuart,J.M., Chang,K., Creighton,C.J. *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
31. Fehlmann,T., Backes,C., Alles,J., Fischer,U., Hart,M., Kern,F., Langseth,H., Rounge,T., Umu,S.U., Kahraman,M. *et al.* (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, **34**, 1621–1628.
32. Langmead,B. (2010) Aligning short sequencing reads with bowtie. *Curr. Protoc. Bioinformatics*, doi:10.1002/0471250953.bi1107s32.
33. Ludwig,N., Fehlmann,T., Kern,F., Gogol,M., Maetzler,W., Deutscher,S., Gurlit,S., Schulte,C., von Thaler,A.K., Deuschle,C. *et al.* (2019) Machine learning to detect Alzheimer's disease from circulating non-coding RNAs. *Genomics Proteomics Bioinformatics*, **17**, 430–440.
34. Leidinger,P., Backes,C., Deutscher,S., Schmitt,K., Mueller,S.C., Frese,K., Haas,J., Ruprecht,K., Paul,F., Stähler,C. *et al.* (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.*, **14**, R78.
35. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimer's Dementia*, **12**, 565–576.
36. Pearson,K. (1901) LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, **2**, 559–572.
37. Leung,C.M., Li,S.C., Chen,T.W., Ho,M.R., Hu,L.Y., Liu,W.S., Wu,T.T., Hsu,P.C., Chang,H.T. and Tsai,K.W. (2014) Comprehensive microRNA profiling of prostate cancer cells after ionizing radiation treatment. *Oncol. Rep.*, **31**, 1067–1078.
38. Kang,S.M., Choi,J.W., Hong,S.H. and Lee,H.J. (2013) Up-regulation of microRNA* strands by their target transcripts. *Int. J. Mol. Sci.*, **14**, 13231–13240.
39. Starega-Roslan,J., Galka-Marciniak,P. and Krzyzosiak,W.J. (2015) Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer. *Nucleic Acids Res.*, **43**, 10939–10951.
40. Fehlmann,T., Reinheimer,S., Geng,C., Su,X., Drmanac,S., Alexeev,A., Zhang,C., Backes,C., Ludwig,N., Hart,M. *et al.* (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet.*, **8**, 123.
41. Ludwig,N., Becker,M., Schumann,T., Speer,T., Fehlmann,T., Keller,A. and Meese,E. (2017) Bias in recent miRBase annotations potentially associated with RNA quality issues. *Sci. Rep.*, **7**, 5162.
42. Backes,C., Sedaghat-Hamedani,F., Frese,K., Hart,M., Ludwig,N., Meder,B., Meese,E. and Keller,A. (2016) Bias in High-Throughput Analysis of miRNAs and Implications for Biomarker Studies. *Anal. Chem.*, **88**, 2088–2095.
43. Kuo,W.T., Ho,M.R., Wu,C.W., Fang,W.L., Huang,K.H. and Lin,W.C. (2015) Interrogation of MicroRNAs involved in gastric cancer using 5p-arm and 3p-arm annotated MicroRNAs. *Anticancer Res.*, **35**, 1345–1352.
44. Jackson,T.J., Spriggs,R.V., Burgoyne,N.J., Jones,C. and Willis,A.E. (2014) Evaluating bias-reducing protocols for RNA sequencing library preparation. *BMC Genomics*, **15**, 569.

3.4 A high-resolution map of the human small non-coding transcriptome

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

This article is available under: https://doi.org/10.1093/bioinformatics/btx814

D160–D167 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 25 September 2017*
doi: 10.1093/nar/gkx851

miRCarta: a central repository for collecting miRNA candidates

Christina Backes[1,*], Tobias Fehlmann[1], Fabian Kern[1], Tim Kehl[2], Hans-Peter Lenhof[2], Eckart Meese[3] and Andreas Keller[1]

[1]Chair for Clinical Bioinformatics, Saarland Informatics Campus, Saarland University, Germany, [2]Center for Bioinformatics, Saarland Informatics Campus, Saarland University, Germany and [3]Institute for Human Genetics, Medical School, Saarland University, Germany

Received August 11, 2017; Revised September 11, 2017; Editorial Decision September 12, 2017; Accepted September 15, 2017

ABSTRACT

The continuous increase of available biological data as consequence of modern high-throughput technologies poses new challenges for analysis techniques and database applications. Especially for miRNAs, one class of small non-coding RNAs, many algorithms have been developed to predict new candidates from next-generation sequencing data. While the amount of publications describing novel miRNA candidates keeps steadily increasing, the current gold standard database for miRNAs - miRBase - has not been updated since June 2014. As a result, publications describing new miRNA candidates in the last three to five years might have a substantial overlap of candidates without noticing. With miRCarta we implemented a database to collect novel miRNA candidates and augment the information provided by miRBase. In the first stage, miRCarta is thought to be a highly sensitive collection of potential miRNA candidates with a high degree of analysis functionality, annotations and details on each miRNA. We added—besides the full content of the miRBase—12,857 human miRNA precursors to miRCarta. Users can match their own predictions to the entries of miRCarta to reduce potential redundancies in their studies. miRCarta provides the most comprehensive collection of human miRNAs and miRNA candidates to form a basis for further refinement and validation studies. The database is freely accessible at https://mircarta.cs.uni-saarland.de/.

INTRODUCTION

MicroRNAs (miRNAs) play a central role in post-transcriptional gene regulation. This class of short non-coding RNAs with an average length of 17–23 nucleotides can bind to their complementary target mRNAs and repress their translation or mediate their degradation (1–3). Since one miRNA potentially regulates many genes and may therefore severely influence the overall regulation network, their expression changes have been the focus of many publications describing various diseases (4–8) and are discussed as potential biomarkers (9–13).

The central repository for miRNAs is the miRBase database (14), which is currently at its 21st version (released 06/14). The last update of miRBase has been >3 years ago. This is problematic for several reasons. Firstly, many miRNA prediction algorithms have been developed and applied to next-generation sequencing (NGS) data in recent years. The published results of these predictions often claim to have found hundreds or thousands of new miRNA candidates (15–18). Since these candidates were so far not integrated in miRBase, different studies contain substantial redundancies. Secondly, several independent groups have found that the current version of miRBase seems to already contain artifacts, wrongly annotated and false positive miRNAs, probably due to the integration of predicted candidates that were not experimentally validated (16,19–21). To overcome this, the miRBase provides a high-confidence miRNA set defined by lower thresholds of reads that must be mapping to the mature sequences besides other rules. In their latest publication, they collated 305 deep sequencing data sets from 38 species to annotate these high-confidence miRNA sets (14). In the meantime, the publicly available small RNA sequencing data has increased exponentially and should be used to further refine a current high-confidence miRNA set. In addition, even some validated miRNAs have not yet made their way into miRBase.

With miRCarta, we aimed to develop a database to bridge the gap between the already available annotations in miRBase and the more recent miRNA predictions from publications or our tool miRMaster (22). To this end, we initially integrated the content of miRBase releases 1.1-21 and enhanced our database with new analysis tools, annota-

[*]To whom correspondence should be addressed. Tel: +49 681 302 68607; Email: c.backes@mx.uni-saarland.de

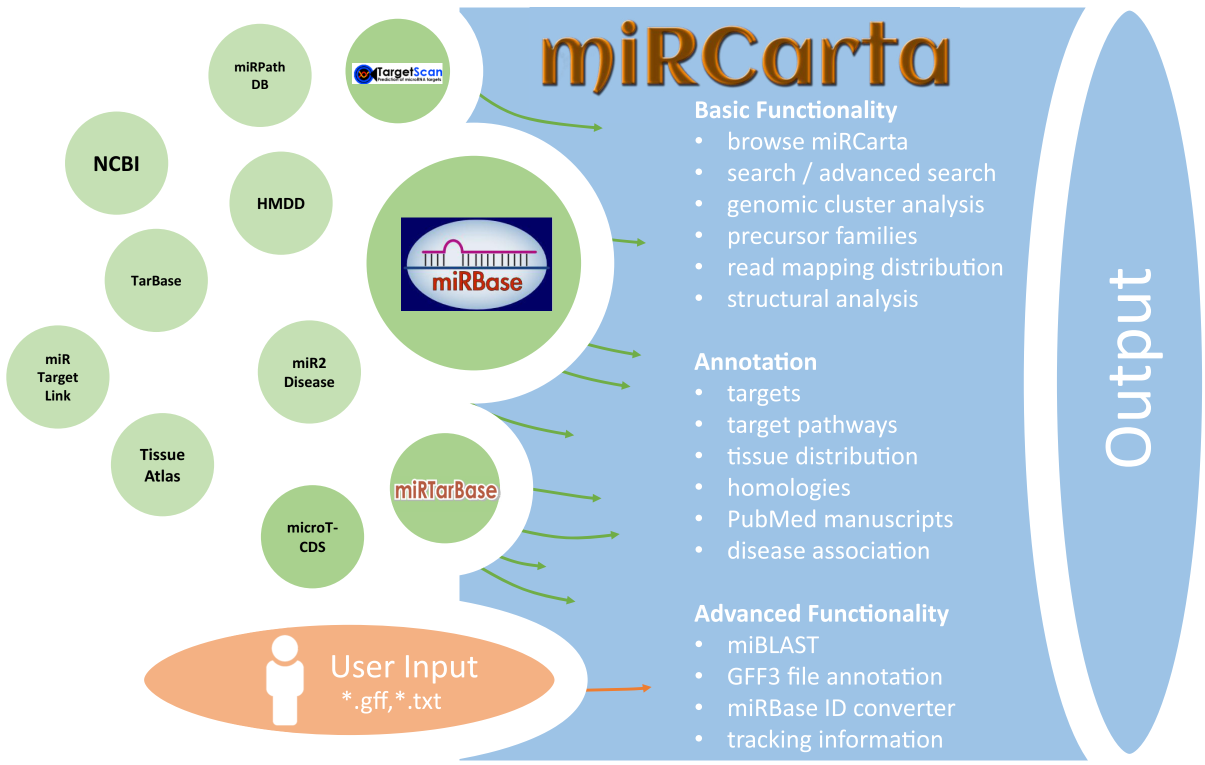


Figure 1. Overview of the integrated or linked data sources and the functionality of miRCarta.

tions, and background information on miRNAs. This part of miRCarta can be used as if querying miRBase and is independent of the remaining updated database content. In a next step, we retrieved updated genomes for 148 organisms that had miRNA annotations in miRBase and remapped the miRNAs to get up-to-date locations for these organisms. We put our focus on the most frequently studied organism in biomedical research, namely *Homo sapiens*. For human, we collected over 18 000 small RNA sequencing data sets from the Sequence Read Archive (SRA) (23), The Cancer Genome Atlas (TCGA) (24) and in-house data sets. These data were processed with our tool miRMaster to predict novel miRNA candidates. To these predictions, we added miRNA candidates from publications (15,16) and miRBase resulting in a total of 24 148 human mature miRNA candidates. To facilitate the decision if the integrated candidates are potentially real miRNAs, we visualize the expression profiles along the corresponding precursors using the mapping results of the 18 035 samples against the stem loop sequences. This way, researchers are able to select conveniently promising candidates for further experimental validation. In addition, we provide a batch query for researchers to match their own miRNA predictions to the entries of miRCarta to reduce potential redundancies in their studies.

This first release of miRCarta provides the most comprehensive collection of human miRNA candidates to date and can serve as an entry point for researchers working in this field searching for current miRNA annotations and predictions. miRCarta is freely accessible at: https://mircarta.cs.uni-saarland.de/.

DATA SOURCES

miRCarta was conceived to provide the information of miRBase (14) as well as more recent data stemming from miRNA predictions. To this end, we integrated the content of miRBase releases 1.1-21, including all naming and sequence changes that the entries of miRBase underwent, as well as the location information for miRNAs in the latest miRBase release comprising 108 organisms. We enhanced this basic information by adding additional data sources and links to external databases. To be able to also filter for miRNA targets, we integrated miRNA target predictions from microT-CDS v5.0 (25) and TargetScan v7.1 (26) and the experimentally validated targets from miRTarBase v6.1 (27). For miRBase precursors, we added links to the Human microRNA Disease Database (HMDD) (28) and to NCBI Gene if official gene symbols were available. miRBase miRNAs are linked to the miRNA pathway dictionary (miRPathDB) (29), miRTargetLink Human (30), Tissue Atlas (31), miR2Disease (32) and TarBase (33). Since

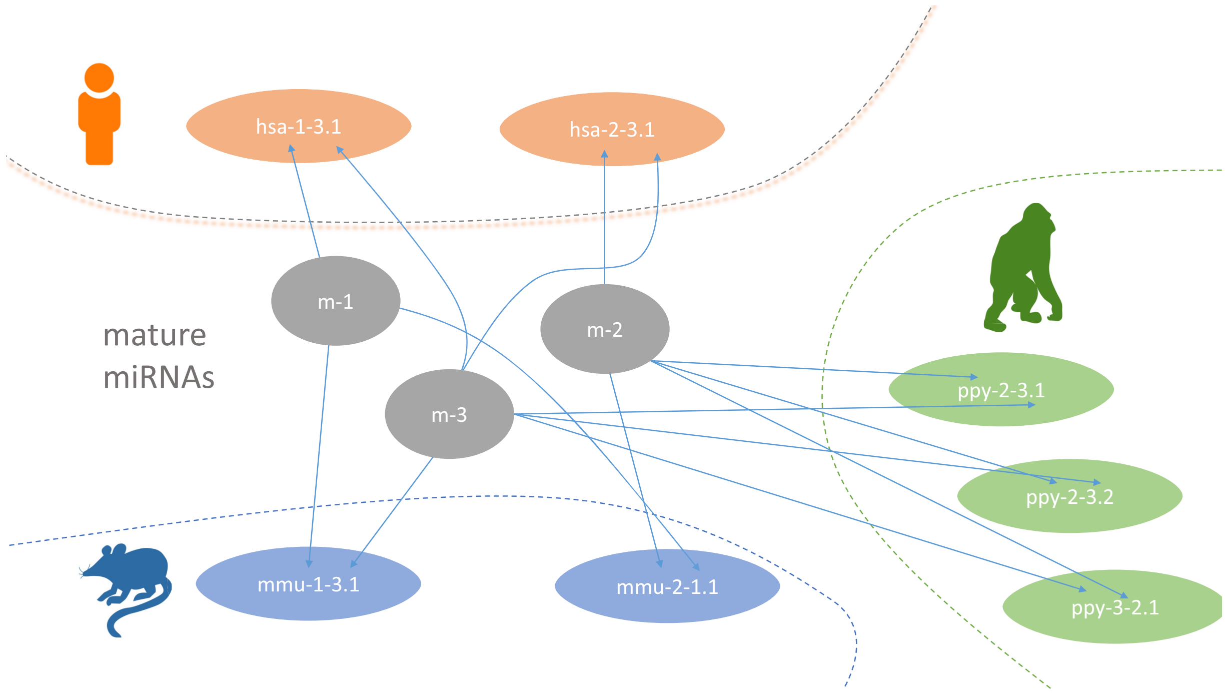


Figure 2. Figurative example for the new naming scheme in miRCarta. MiRNAs are named with m-[number] and are organism unspecific. Precursors are named [organism_abbreviation]-[5p miRNA]-[3p miRNA].[location ID]. In this example, we have a human precursor hsa-1-3.1, consisting of miRNAs m-1 and m-3. If this precursor has another location in the genome it gets another location ID as exemplified for ppy-2-3.1 and ppy-2-3.2.

the downloaded and integrated data of miRTarBase and microT-CDS are slightly different from their online versions, links to their web sites were also added. Obviously, all miRBase entries integrated in miRCarta are also connected to their original source in miRBase. An overview of miRCarta's data sources, external links and functionality is illustrated in Figure 1.

RE-ANNOTATING miRNAs/PRECURSORS AND NEW NAMING SCHEME

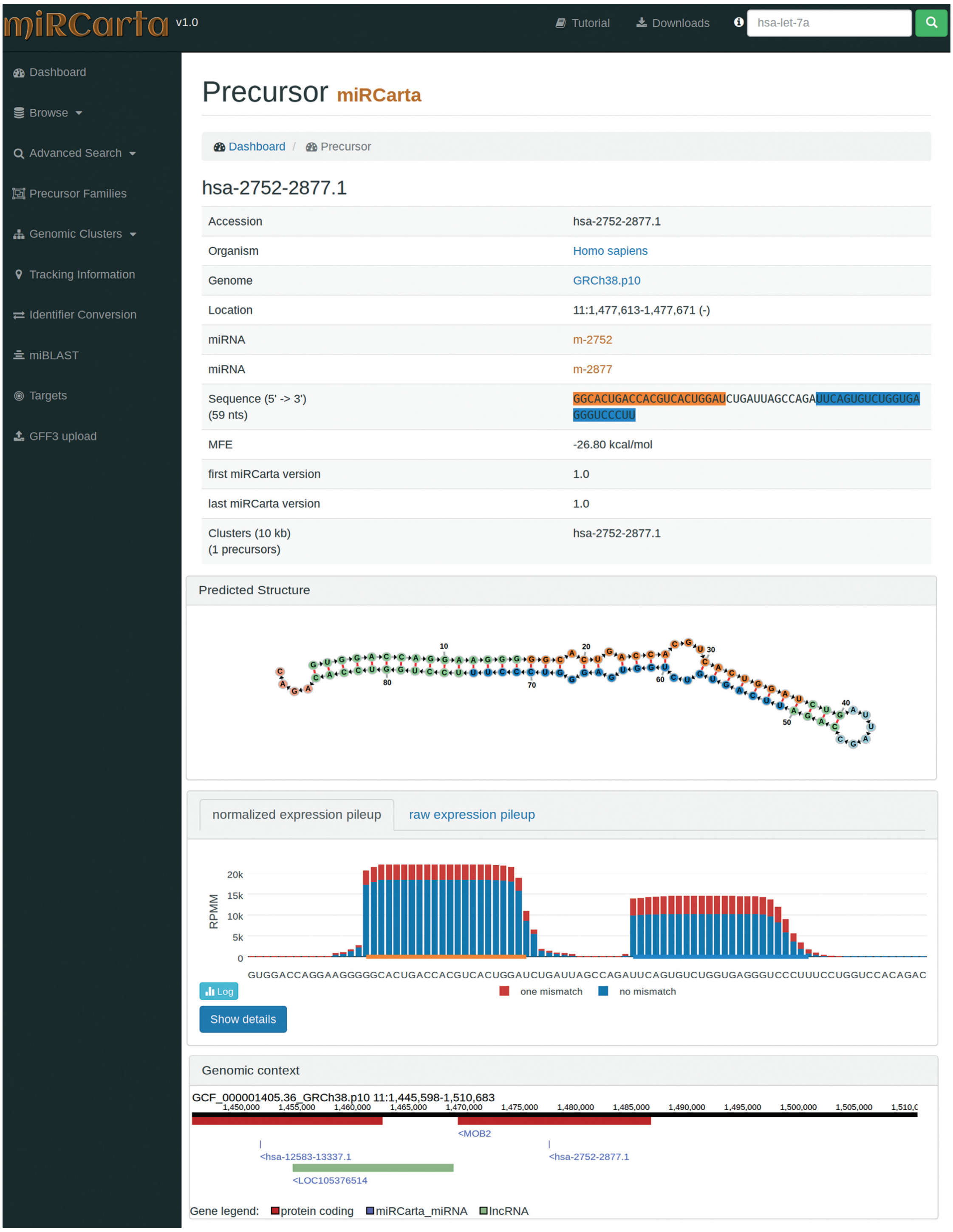
We collected the newest genome releases from NCBI RefSeq/Genbank (34) for 148 organisms that had miRNA/precursor annotations in miRBase. In brief, the precursors from miRBase for these 148 organisms were mapped with Bowtie 1.1.2 (35) against their respective genomes. Since a central aim of miRCarta was to include new potential candidates, a new naming scheme for miRNAs and precursors had to be created (Figure 2). Mature miRNAs in miRCarta are named with m-[number] and are organism unspecific. Precursors are however organism specific, starting with a three letter code for an organism, followed by the number of the 5′ miRNA and the number of the 3′ miRNA and ending with a locus identifier, e.g. hsa-1-52.1 consists of mature miRNAs m-1 and m-52. To improve the miRNA annotations in miRBase, we collected for human 18 035 small RNA sequencing samples from SRA (23), TCGA (24) and in-house data sets. We mapped the reads against the human precursor sequences and derived the sequence of the canonical forms from the expression profiles. Thereby, we processed the miRNAs in their median RPMMM (reads per million mapped to miRNAs) expression order across all our samples, resulting in the most expressed miRNA as m-1 (corresponding to hsa-miR-21-5p) and so on. Using these sequences as basis, we added predictions from publications (15,16) and our tool miRMaster (22) for the 18 035 samples to this pool, as well as the re-mapped sequences for the remaining 147 organisms to complete the information added to miRCarta. More details on this integration and naming process can be found in the Supplemental Material.

A side effect of the re-annotation is that a miRNA in miRBase might not be identical to a miRNA in miRCarta anymore, e.g. it can be shifted to the left or right or have a different length. Still, we deemed this re-annotation necessary since our analyses showed that the currently annotated canonical form represented only in 42% of cases the actually most expressed form across all of our samples. In the web interface we provide links between these miRCarta precursors/miRNAs and miRBase precursors/miRNAs to allow for an easier comparison.

DATABASE IMPLEMENTATION AND FUNCTIONALITY

Implementation

miRCarta consists of a MySQL database and a MongoDB NoSQL database. The MySQL part contains organisms, sequences, locations, miRNAs, precursors and targets, while the NoSQL database stores the expression data as matrix



Precursor miRCarta

Dashboard / Precursor

hsa-2752-2877.1

Accession	hsa-2752-2877.1
Organism	Homo sapiens
Genome	GRCh38.p10
Location	11:1,477,613-1,477,671 (-)
miRNA	m-2752
miRNA	m-2877
Sequence (5' -> 3') (59 nts)	GGCACUGACCACGUCACUGGAUCUGAUUAGCCAGAUUCAGUGUCUGGUGA GGGUCCCUU
MFE	-26.80 kcal/mol
first miRCarta version	1.0
last miRCarta version	1.0
Clusters (10 kb) (1 precursors)	hsa-2752-2877.1

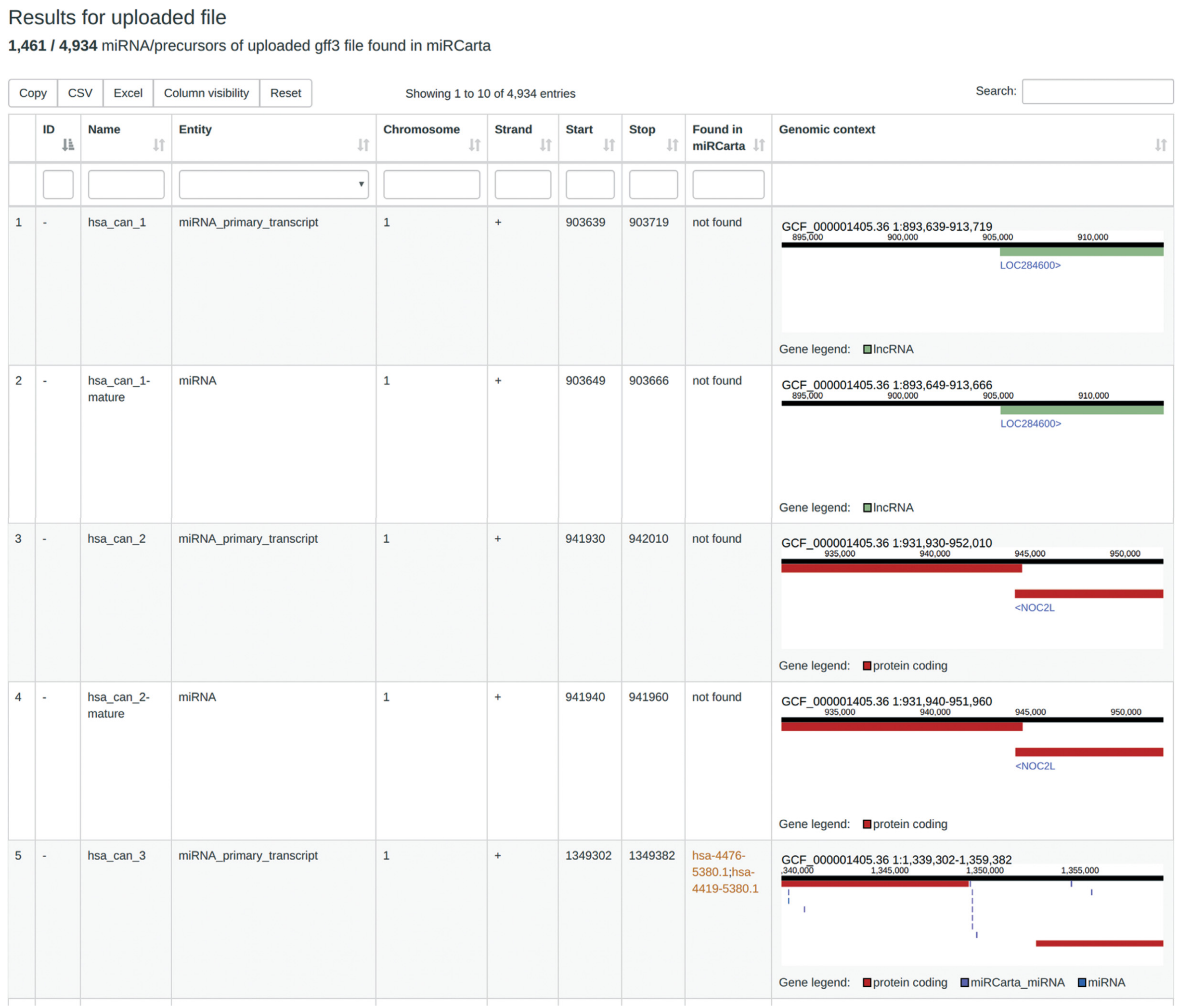
Predicted Structure

normalized expression pileup | raw expression pileup

RPMM
20k
15k
10k
5k
0
GUGGACCAGGAAGGGGGCACUGACCACGUCACUGGAUCUGAUUAGCCAGAUUCAGUGUCUGGUGAGGGUCCCUUUCCUGGUCCACAGAC
Log
🟥 one mismatch
🟦 no mismatch
Show details

Genomic context

Figure 3. Example of a precursor view for a predicted candidate in miRCarta. First, we list several basic facts about the precursor like its sequence, location, links to miRNAs, etc. In addition, we visualize the stem loop structure with the FornaContainer plugin (36) and color the miRNAs in the same way as in the sequence of the precursor. Below the structure, we show the pileup plots for the normalized or raw read counts with plotly.js. The user can easily switch here between log and linear scale or even visualize only counts with zero or one mismatches. The button 'Show details' opens a new HTML page, where more information can be found on how many samples had reads for this precursor and graphics showing if we found these reads rather continuously in several experiments or only a few. The last part shows the genomic context of the current precursor in a window of ±10 kb. This way it can be easily assessed if there are more precursors in this range or if the precursor lies in a gene or close to a gene for example. The genomic context is also interactive and allows for zooming in and out, and shows more information when clicking on a gene or miRNA etc.



Results for uploaded file

1,461 / 4,934 miRNA/precursors of uploaded gff3 file found in miRCarta

	ID	Name	Entity	Chromosome	Strand	Start	Stop	Found in miRCarta	Genomic context
1	-	hsa_can_1	miRNA_primary_transcript	1	+	903639	903719	not found	GCF_000001405.36 1:893,639-913,719 895,000 900,000 905,000 910,000 LOC284600> Gene legend: 🟩lncRNA
2	-	hsa_can_1-mature	miRNA	1	+	903649	903666	not found	GCF_000001405.36 1:893,649-913,666 895,000 900,000 905,000 910,000 LOC284600> Gene legend: 🟩lncRNA
3	-	hsa_can_2	miRNA_primary_transcript	1	+	941930	942010	not found	GCF_000001405.36 1:931,930-952,010 935,000 940,000 945,000 950,000 <NOC2L Gene legend: 🟥protein coding
4	-	hsa_can_2-mature	miRNA	1	+	941940	941960	not found	GCF_000001405.36 1:931,940-951,960 935,000 940,000 945,000 950,000 <NOC2L Gene legend: 🟥protein coding
5	-	hsa_can_3	miRNA_primary_transcript	1	+	1349302	1349382	hsa-4476-5380.1;hsa-4419-5380.1	GCF_000001405.36 1:1,339,302-1,359,382 ,340,000 1,345,000 1,350,000 1,355,000 Gene legend: 🟥protein coding 🟪miRCarta_miRNA 🟦miRNA

Figure 4. Excerpt of the results of uploading a GFF3 file for the predictions of Friedländer *et al.* (40). We find overlaps for 1461 of 4934 uploaded precursors/miRNAs in miRCarta. The first four rows in the table show examples for entries we did not find in miRCarta and the genomic context view shows that there are also no other miRNAs in a window of ±10 kb around the annotated location. The fifth row shows an entry where we have overlaps in miRCarta and the genomic context view illustrates that there are many other miRNA annotations in range.

format for a more efficient access. The MySQL database schema is illustrated in Supplemental Figure S1. The server-side backend of the web application uses Django 1.11 and is written in Python 3. The web interface is implemented in Django's HTML template language and is enhanced with several JavaScript libraries for a more interactive user experience. The tables we visualize are created with the jQuery plugin DataTables, the genomic context visualization is done using TnT Genome, and the expression profile plots are rendered with plotly.js, the structure visualization with FornaContainer (36). For styling we use Bootstrap 3 and custom CSS files.

Functionality

miRCarta integrates the miRBase database and additional new miRNA candidates, expression data, updated organisms, and genomes as entry point for miRNA researchers. As illustrated in Figure 1, miRCarta provides different levels of functionality.

Basic functionality

Browse. The classical entry point 'Browse' lists all miRNAs and precursors for a selected organism. For precursors, this view also visualizes the normalized read counts of the mapped NGS data without and with mismatches. This way a user can assess if the expression profile over a (putative or known) precursor seems likely for miRNA expression and more rapidly identify real precursors/miRNAs from false positive annotations.

Advanced search. Using 'Advanced Search', users can restrict their query results to certain miRNAs and/or precursors of certain organisms that might have been validated with a certain experiment and so on. The results are visualized as HTML table, unless one of the download options is checked.

Precursor families. For the miRBase content, we also integrated the precursor families. A user can search for precursor names or miRBase accession numbers or select an organism and get all precursor families of the input as result. If nothing is selected all precursor families will be listed.

Genomic clusters. 'Genomic Clusters' visualizes clusters of precursors within a selectable window size in a tabular format and as stacked bar plots along the chromosomes. In Supplemental Figure S2, we queried the miRBase part for clusters in *Homo sapiens*. The stacked bar plot visualization directly shows that the largest clusters can be found on chromosomes 14 and 19.

Read mapping distribution. For human precursors, we visualize the mappings of the sequencing reads of the collected 18 035 samples with and without mismatch. The pileup plots can be found in the single precursor views (Figure 3) as well as in the tabular overviews of 'Browse' and 'Advanced Search' for *H. sapiens*. In the precursor view the pileup can be switched between normalized and raw read counts, log and linear scale, and also visualize perfect matching reads and reads with one mismatch separately or combined. More information about the number of mapped reads and the corresponding number of different samples for the precursor can be found on a separate HTML page by clicking on 'Show details' below the plot.

Structural analysis. The secondary structures for the precursor sequences are computed with RNAfold (37) and visualized with FornaContainer (36). This illustration is available in the precursor specific views.

Annotation

Targets. Since we integrated miRTarBase, microT-CDS, and TargetScan, miRCarta can provide a combined search of miRNAs and targets using experimentally validated or predicted targets, respectively. If all three databases are selected, the resulting table will contain for each database a column with either 0 or 1 as entry, which can be used for sorting and filtering for results that have e.g. only hits in all three target databases.

Target pathways. For potential target pathways, we linked the tools MiRTargetLink (30) and miRPathDB (29) for miRNAs. The links can be found on the right hand side of the miRNA views if they are available.

Tissue distribution. For miRNAs we also provide links to the tool TissueAtlas (31), which shows the miRNA abundance in 61 tissue biopsies of two individuals.

Homologies. We mapped the miRCarta miRNAs with their respective flanks (see Supplemental Material) against all 148 organisms without mismatch. If such a miRNA sequence is found in an organism where it has not been annotated so far, we list these findings under 'miRNA homologies' on the right hand side of the miRNA view.

PubMed manuscripts. We provide links to the manuscripts describing miRNAs/precursors in miRBase, as well as validation experiments for the targets in miRTarbase.

Disease association. For disease association, we provide links for miRNAs to miR2Disease (32) and to HMDD (28) for precursors.

Advanced functionality

miBLAST. Under 'miBLAST', users can enter a miRNA sequence and get the BLAST (38) results for miRBase and miRCarta miRNAs.

GFF3 file annotation. Using the analysis tool 'GFF3 upload' users can upload their own standard GFF3 files containing e.g. the locations of predictions of miRNAs and precursors for a certain organism. The data is matched against the available miRCarta entries and the result is visualized as a table, which shows how many findings are new or have overlaps with entries in miRCarta.

miRBase ID converter. The naming of miRNAs changed during different releases of miRBase, which can cause problems when comparing findings to older manuscripts where a different miRBase release was used. With the 'Identifier Conversion' tool researchers can convert their miRBase names into the latest available version.

Tracking Information. Inspired by the tool miRBase Tracker (39), we provide tracking information for each identifier in miRBase, which allows in a straightforward way to illustrate the changes a miRBase name or sequence underwent during different miRBase releases.

Application examples

To demonstrate the 'GFF3 Upload' functionality, we collected the predictions from Friedländer *et al.* (40) as independent test set, converted the locations with liftOver into GRCh38 coordinates and created a gff3 file. This file was uploaded in miRCarta using the default parameters. In Figure 4, we can see the first five entries of which four have not been found in miRCarta. The fifth has an overlap with two entries in miRCarta. The genomic context visualization is especially helpful if no overlap has been found to assess whether other miRNA precursors might be in range. Altogether, 1461 of 4934 entries from the Friedländer dataset have already been annotated in miRCarta.

miRCarta's precursor view enables users to grasp the structure and expression profiles more easily than it was possible using miRBase. In Supplemental Figure 3 we visualize the precursor hsa-mir-5739 in miRBase on the left-hand side and in miRCarta on the right-hand side. The structure in miRBase is visualized as ASCII code and it is hard to assess for this precursor if this is a good structure or not. In the miRCarta view, it is directly clear that this is not a valid precursor by looking at the folding structure and the expression profile below.

Since we visualize the expression profiles also in 'Browse' and 'Advanced Search' for precursors in *H. sapiens*, these plots can also be easily used to scroll through longer lists of precursors and select interesting candidates for further validations.

FUTURE WORK

While still undiscovered miRNAs may exist, the current collection of miRCarta represents a substantial part of the human miRNome. This set—tailored to be a very sensitive collection of miRNAs—contains certainly a large number of false positive predictions. Nevertheless, this set will be useful for other researchers to match their own predictions against it to reduce redundancies in their studies. This current 'high-sensitivity' set will form the basis for our further developments. In a next step, we will reduce the 'high-sensitivity' set by merging similar findings, e.g. overlapping precursors from different predictions. This will result in a collapsed set with slightly lesser sensitivity than the original set. At last, we will create a 'high-specificity' set which will consist of experimentally validated miRNAs relying on cloning the precursor and providing evidence for the mature miRNAs using Northern Blots. In addition, we will also annotate miRNA isoforms and include more information about other organisms.

CONCLUSION

miRCarta bridges the gap between established annotations in miRBase and more recent miRNA predictions from publications and our software miRMaster. In our proof-of-concept study for human we succeeded to demonstrate that these candidates—which certainly contain false positive hits—contain interesting candidates for further validation. With this approach, we aim to create a high-resolution map of potential human miRNAs, which will be refined in further releases to create finally a set of experimentally validated real miRNAs.

DATA AVAILABILITY

miRCarta is publicly accessible at https://mircarta.cs.uni-saarland.de/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

Funding for open access charge: Michael J. Fox Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Sontheimer,E.J. and Carthew,R.W. (2005) Silence from within: endogenous siRNAs and miRNAs. *Cell*, **122**, 9–12.
3. Filipowicz,W., Jaskiewicz,L., Kolb,F.A. and Pillai,R.S. (2005) Post-transcriptional gene silencing by siRNAs and miRNAs. *Curr. Opin. Struct. Biol.*, **15**, 331–341.
4. Li,Y.J., Ping,C., Tang,J. and Zhang,W. (2016) MicroRNA-455 suppresses non-small cell lung cancer through targeting ZEB1. *Cell Biol. Int.*, **40**, 621–628.
5. Keller,A., Leidinger,P., Steinmeyer,F., Stahler,C., Franke,A., Hemmrich-Stanisak,G., Kappel,A., Wright,I., Dorr,J., Paul,F. *et al.* (2014) Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. *Mult. Scler.*, **20**, 295–303.
6. Keller,A., Leidinger,P., Vogel,B., Backes,C., ElSharawy,A., Galata,V., Müller,S., Marquart,S., Schrauder,M., Strick,R. *et al.* (2014) miRNAs can be generally associated with human pathologies as exemplified for miR-144*. *BMC Med.*, **12**, 224.
7. Fenoglio,C., Ridolfi,E., Cantoni,C., De Riz,M., Bonsi,R., Serpente,M., Villa,C., Pietroboni,A.M., Naismith,R.T., Alvarez,E. *et al.* (2013) Decreased circulating miRNA levels in patients with primary progressive multiple sclerosis. *Mult. Scler.*, **19**, 1938–1942.
8. Keller,A., Leidinger,P., Bauer,A., Elsharawy,A., Haas,J., Backes,C., Wendschlag,A., Giese,N., Tjaden,C., Ott,K. *et al.* (2011) Toward the blood-borne miRNome of human diseases. *Nat. Methods*, **8**, 841–843.
9. Roth,P., Keller,A., Hoheisel,J.D., Codo,P., Bauer,A.S., Backes,C., Leidinger,P., Meese,E., Thiel,E., Korfel,A. *et al.* (2015) Differentially regulated miRNAs as prognostic biomarkers in the blood of primary CNS lymphoma patients. *Eur. J. Cancer*, **51**, 382–390.
10. Ma,J., Lin,Y., Zhan,M., Mann,D.L., Stass,S.A. and Jiang,F. (2015) Differential miRNA expressions in peripheral blood mononuclear cells for diagnosis of lung cancer. *Lab. Invest.*, **95**, 1197–1206.
11. Sayed,A.S., Xia,K., Yang,T.L. and Peng,J. (2013) Circulating microRNAs: a potential role in diagnosis and prognosis of acute myocardial infarction. *Dis. Markers*, **35**, 561–566.
12. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement.*, **12**, 565–576.
13. Margue,C., Reinsbach,S., Philippidou,D., Beaume,N., Walters,C., Schneider,J.G., Nashan,D., Behrmann,I. and Kreis,S. (2015) Comparison of a healthy miRNome with melanoma patient miRNomes: are microRNAs suitable serum biomarkers for cancer? *Oncotarget*, **6**, 12110–12127.
14. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
15. Londin,E., Loher,P., Telonis,A.G., Quann,K., Clark,P., Jing,Y., Hatzimichael,E., Kirino,Y., Honda,S., Lally,M. *et al.* (2015) Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E1106–E1115.
16. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grasser,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
17. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
18. Jha,A., Panzade,G., Pandey,R. and Shankar,R. (2015) A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Res.*, **43**, 8713–8724.
19. Vitsios,D.M., Davis,M.P., van Dongen,S. and Enright,A.J. (2016) Large-scale analysis of microRNA expression, epi-transcriptomic features and biogenesis. *Nucleic Acids Res.*, **45**, 1079–1090.
20. Fromm,B., Billipp,T., Peck,L.E., Johansen,M., Tarver,J.E., King,B.L., Newcomb,J.M., Sempere,L.F., Flatmark,K., Hovig,E. *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.*, **49**, 213–242.

21. Ludwig,N., Becker,M., Schumann,T., Speer,T., Fehlmann,T., Keller,A. and Meese,E. (2017) Bias in recent miRBase annotations potentially associated with RNA quality issues. *Sci. Rep.*, **7**, 5162.
22. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A., Würstle,M., Hübenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.
23. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database, C. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
24. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
25. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Vlachos,I.S., Vergoulis,T., Reczko,M., Filippidis,C., Dalamagas,T. and Hatzigeorgiou,A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, **41**, W169–W173.
26. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, doi:10.7554/eLife.05005.
27. Chou,C.H., Chang,N.W., Shrestha,S., Hsu,S.D., Lin,Y.L., Lee,W.H., Yang,C.D., Hong,H.C., Wei,T.Y., Tu,S.J. *et al.* (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.
28. Li,Y., Qiu,C., Tu,J., Geng,B., Yang,J., Jiang,T. and Cui,Q. (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.*, **42**, D1070–D1074.
29. Backes,C., Kehl,T., Stockel,D., Fehlmann,T., Schneider,L., Meese,E., Lenhof,H.P. and Keller,A. (2017) miRPathDB: a new dictionary on microRNAs and target pathways. *Nucleic Acids Res.*, **45**, D90–D96.
30. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) MiRTargetLink–miRNAs, genes and interaction networks. *Int. J. Mol. Sci.*, **17**, 564.
31. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stahler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
32. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
33. Vlachos,I.S., Paraskevopoulou,M.D., Karagkouni,D., Georgakilas,G., Vergoulis,T., Kanellos,I., Anastasopoulos,I.L., Maniou,S., Karathanou,K., Kalfakakou,D. *et al.* (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.*, **43**, D153–D159.
34. Coordinators,N.R. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
35. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
36. Kerpedjiev,P., Hammer,S. and Hofacker,I.L. (2015) Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**, 3377–3379.
37. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
38. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
39. Van Peer,G., Lefever,S., Anckaert,J., Beckers,A., Rihani,A., Van Goethem,A., Volders,P.J., Zeka,F., Ongenaert,M., Mestdagh,P. *et al.* (2014) miRBase Tracker: keeping track of microRNA annotation changes. *Database (Oxford)*, **2014**, bau080.
40. Friedlander,M.R., Lizano,E., Houben,A.J., Bezdan,D., Banez-Coronel,M., Kudla,G., Mateu-Huertas,E., Kagerbauer,B., Gonzalez,J., Chen,K.C. *et al.* (2014) Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol.*, **15**, R57.

3.6 Large-scale validation of miRNAs by disease association, evolutionary conservation and pathway activity

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

This article is available under: https://doi.org/10.1080/15476286.2018.1559689

3.7 *An estimate of the total number of true human miRNAs*

Published online 1 March 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 7* ***3353–3364***
doi: 10.1093/nar/gkz097

An estimate of the total number of true human miRNAs

Julia Alles[1,*,†], Tobias Fehlmann[2,†], Ulrike Fischer[1], Christina Backes[2], Valentina Galata[2], Marie Minet[1,2], Martin Hart[1], Masood Abu-Halima[1], Friedrich A. Grässer[3], Hans-Peter Lenhof[4], Andreas Keller[2,*,†] and Eckart Meese[1,†]

[1]Institute of Human Genetics, Saarland University, 66421 Homburg, Germany, [2]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [3]Institute of Virology, Saarland University Medical School, 66421 Homburg, Germany and [4]Chair for Bioinformatics, Center for Bioinformatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany

Received December 13, 2018; Revised January 30, 2019; Editorial Decision February 01, 2019; Accepted February 07, 2019

ABSTRACT

While the number of human miRNA candidates continuously increases, only a few of them are completely characterized and experimentally validated. Toward determining the total number of true miRNAs, we employed a combined *in silico* high- and experimental low-throughput validation strategy. We collected 28 866 human small RNA sequencing data sets containing 363.7 billion sequencing reads and excluded falsely annotated and low quality data. Our high-throughput analysis identified 65% of 24 127 mature miRNA candidates as likely false-positives. Using northern blotting, we experimentally validated miRBase entries and novel miRNA candidates. By exogenous overexpression of 108 precursors that encode 205 mature miRNAs, we confirmed 68.5% of the miRBase entries with the confirmation rate going up to 94.4% for the high-confidence entries and 18.3% of the novel miRNA candidates. Analyzing endogenous miRNAs, we verified the expression of 8 miRNAs in 12 different human cell lines. In total, we extrapolated 2300 true human mature miRNAs, 1115 of which are currently annotated in miRBase V22. The experimentally validated miRNAs will contribute to revising targetomes hypothesized by utilizing falsely annotated miRNAs.

INTRODUCTION

MicroRNAs have a major regulatory impact on gene expression by facilitating sequence-specific RNA interference. Mediated by Argonaute proteins and other components of RISC (RNA-induced Silencing Complex), miRNAs bind complementary sequences within mRNA transcripts, which results in decreased expression levels of target proteins (1–3). Variations in miRNA levels have been reported for patients' solid tissues, blood and other body fluids, making miRNAs promising candidates for markers in a manifold of diseases (4–15). The still increasing information density on miRNAs necessitates collecting sequences and annotations in public databases. The reference repository miRBase, currently holds information about 1917 human precursors and 2656 mature miRNAs (release 22) (16).

Since the start of the miRNA registry that later developed into miRBase (17), the number of deposited miRNAs has constantly increased mostly due to high-throughput sequencing of small RNAs. The challenge of the exploding number of miRNAs is an increase of false-positive entries in miRBase and other databases. Since the best possible evidence for each miRNA is among the primary aims of miRBase (18), explicit action was taken since release 5 to reduce the number of false-positives. In 2014, miRBase defined criteria for high-confidence miRNAs, which represented only 16% of the human miRNAs annotated in release 21 (19). In 2018, a new version of miRBase was released, which incorporated additional sequencing data that have been considered to annotate all miRNAs, leading to 26% high-confidence human miRNA annotations (16). Notably, there is a striking drop of high-confidence miRNAs in later versions of miRBase. The increasing number of questionable miRNAs in late miRBase releases apparently results from the aforementioned increasing use of NGS-based approaches (20). This challenge was also reported for mouse miRNAs. Here, nearly a third of the annotations in miRBase version 14.0 was discounted as non-authentic miRNAs (21).

Certainly, there is a need for universal definition of criteria to define true miRNAs (22–25). Consistent naming system and precise *in silico* prediction models can contribute to the quality of miRNA databases. Both highly specific databases (such as miRGeneDB containing a few but high-likely miRNAs (26)) and sensitive databases (such

[*]To whom correspondence should be addressed. Email: ju.al@mx.uni-saarland.de
Correspondence may also be addressed to Andreas Keller. Tel: +49 174 1684638; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

as miRCarta that aims at providing high-likely miRNAs but also a broad collection of published miRNA candidates (27)) are required. However, the quality of respective miRNA databases finally always depends on the availability of highly reliable positive and negative training sets, i.e. miRNAs that have been verified by suitable experimental methods. Like NGS-based approaches, polymerase chain reaction (PCR) also entails an elevated risk of identifying false-positive miRNAs due to its nature as an amplification based approach (28,29). Northern blotting (NB) represents a rather solid technique for the detection of single miRNAs. However, there is an obvious decline of miRNAs detected by NB due to its time-consuming design and its consequently low-throughput character. Only for 3.6% of all human miRNAs listed in miRBase V22, there is evidence of miRNA expression by NB, according to experiments listed in miRBase. Frequently, only endogenous expression of a RNA fragment matching the miRNA sequence is provided (17,20,26,30).

To estimate the total number of human miRNAs, we used a high-throughput *in silico* and a low-throughput experimental approach. First, 28 866 human small RNA sequencing data sets containing 363.7 billion sequencing reads were mapped to the human genome and to 24 127 mature miRNA candidates. After excluding likely false-positives, we selected 108 miRNA precursors giving rise to 205 mature forms, i.e. 84 known and 119 novel candidates to be tested for processing in HEK 293T cells. Endogenous expression of 11 selected miRNAs (5 of which were not tested in HEK 293T) was analyzed in 11 additional human cell lines derived from different tissues including, liver, lung, prostate, bone marrow, cervix, placenta, mammary gland, testis, B- and T-lymphocytes, and keratinocytes. For NB, we used radiolabeled probes designed to detect both the precursors and the according 5p and 3p mature forms. A sketch of the study set-up is presented in Figure 1.

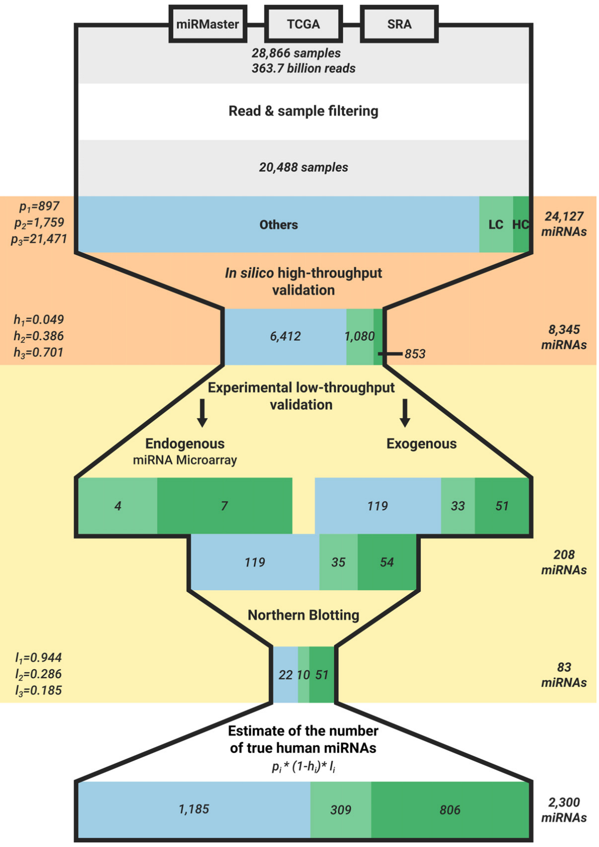


Figure 1. Workflow of the analysis to estimate the number of true human miRNAs. Samples containing NGS data were collected from SRA, TCGA and data uploaded to miRMaster, and the obtained samples and reads were filtered. Three sets of miRNAs were created: miRBase high-confidence (HC), miRBase low-confidence set (LC) and other (Other). Using *in silico* high-throughput validation and experimental low-throughput validation steps, the probabilities that a miRNA will pass the validation procedure have been calculated for each miRNA set respectively. Finally, the number of true miRNAs was estimated using the original miRNA counts and the computed probabilities.

MATERIALS AND METHODS

miRBase-data analysis

All data from miRBase used in this study are available at the miRBase download section (http://www.mirbase.org/ftp.shtml). Changes between two subsequent miRBase releases are collected in miRNA.diff files. miRNA.dead files contain entries that have been removed from the database. Data accumulation and comparisons were carried out both manually and bioinformatically. If not mentioned explicitly, all analyses done in this manuscript are performed on mature miRNAs.

miRCarta/miRMaster analysis

To obtain a collection of as many as possible NGS data sets, we integrated data from miRCarta (27), as well as additional data from the sequence read archive that was added between February and November 2017. A detailed description of the data collection and filtering can be found in our recently published study (31). In brief, the sample collection stems from three different sources, i.e. the sequence read archive (32), the cancer genome atlas (TCGA) and data sets analyzed with miRMaster (33). miRMaster contains data sets from users that applied it for data analysis and volunteered to make their data available in aggregated form for secondary analysis. Altogether, 28 866 human small RNA sequencing data sets containing 363.7 billion reads were integrated. A stringent quality filtering was applied to verify the data integrity. The majority of reads had to match to the human genome (to avoid contamination by other species that erroneously have been annotated as human samples) and of those the majority was not allowed to match to mRNAs (to avoid transcriptome sequencing erroneously annotated as small RNA sequencing and low quality data sets that contain many fragmented mRNAs). From the remaining 20 488 data sets, valid mature miRNAs have been described by applying three criteria on the read profile of their precursors. To this end, we mapped the reads of all data sets against all precursors and allowed no mismatches. We normalized the expression of each read to reads per million mapped to miRNAs to account for different sequencing depth and library composition. First, we determined the 5′ homogeneity of the dominant miRNA and required that over 50% of the normalized reads that mapped to the miRNA start at the same 5′ position. Thereby, we accounted for the fact that miRNAs have a very low 5′ end variability. Second, we considered the reads that did not map in accordance with Dicer processing. Therefore, we determined which fraction of the reads did not map with a variability of two bases at the 5′ end and five bases at the 3′ end to the annotated miRNAs. If this fraction accounted for more than 25% of the normalized reads, we discarded the precursor. Third, we determined the number of valid stacks that could be found by evaluating the coverage profile of the precursors. We defined a stack as the longest stretch of bases for which the most covered base differed from the lowest covered one by at most 20%. A stack was considered valid if it spanned between 16 and 29 bases. We required at least one valid stack per precursor. This criteria accounts for coverage profiles that exhibit clear read stacks, as expected from miRNA precursors. Finally, we kept all miRNAs that resulted from the precursors that fulfilled all criteria.

IsomiR analysis

IsomiR variants for miRNAs of miRBase V22 were determined using miRMaster (33), based on the 20 488 NGS data sets described above. Briefly, reads were mapped to all annotated miRNA precursors while allowing up to two non-template additions at both ends and one mismatch in between. IsomiRs were then determined relative to the coordinates of the annotated miRNAs in miRBase. Reads were counted when their mapping position differed at most two nucleotides at the 5′ end and five nucleotides at the 3′ end. Finally, an isomiR was determined to be present when totaling at least 2% of the total reads per million mapped to miRNA normalized counts.

Construction of miRNA expression vectors

Inserts for miRNA expression vectors with pSG5 backbone (Stratagene, now Agilent Technologies, Santa Clara, California) were synthesized and cloned into pEX-A2 vectors by Eurofins Genomics (Ebersberg, Germany). Therefore, hsa-miRNA precursor sequences were pasted into the UCSC genome browser BLAT tool (GRCh38/hg38 assembly) and correspondent DNA sequences with 100 additional bases up- and downstream and flanking EcoRI/BamHI/BglII restriction sites were used for custom gene synthesis by Eurofins Genomics (Ebersberg, Germany). Lyophilisates were reconstituted at 100 ng/μl with H_2O and 1 μg of DNA was digested using appropriate EcoRI/BamHI/BglII restriction enzymes. Pre-mir inserts were subcloned into pSG5 vector. Positive clones were determined by colony-PCR, restriction digestion and sanger sequencing (Seq-It, Kaiserslautern, Germany and Eurofins Genomics, Ebersberg, Germany). pSG5-miRNA expression vectors that were not synthesized by Eurofins Genomics have been cloned by PCR amplification before.

Cell culture and transfection

A549, HaCaT, HeLa and HUH-7 cells were cultivated at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium (Life Technologies, Darmstadt, Germany) supplemented with 10% fetal bovine serum (Biochrom, Berlin, Germany) and antibiotics (100 U/ml Penicillin, 100 μg/ml Streptomycin) (Life Technologies, Darmstadt, Germany). SHSY-5Y cells were cultivated at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium (Life Technologies, Darmstadt, Germany) supplemented with 20% fetal bovine serum (Biochrom, Berlin, Germany) and antibiotics (100 U/ml Penicillin and 100 μg/ml Streptomycin) (Life Technologies, Darmstadt, Germany). MCF-7, PC-3, DG-75 and Jurkat cells were cultivated at 37°C and 5% CO_2 in RPMI 1640 Medium (Life Technologies, Darmstadt, Germany) supplemented with 10% fetal bovine serum (Biochrom, Berlin, Germany) and antibiotics (100 U/ml Penicillin and 100 μg/ml Streptomycin) (Life Technologies, Darmstadt, Germany). JEG-3 cells were cultivated at 37°C and 5% CO_2 in Ham's F12 Nutrient Mixture Medium (Life Technologies, Darmstadt, Germany) supplemented with 10% fetal bovine serum (Biochrom, Berlin, Germany) and antibiotics (100 U/ml Penicillin and 100 μg/ml Streptomycin) (Life Technologies, Darmstadt, Germany). Tera-1 cells were cultivated at 37°C and 5% CO_2 in McCoy's 5A Medium (Life Technologies, Darmstadt, Germany) supplemented with 15% fetal bovine serum (Biochrom, Berlin, Germany) and antibiotics (100 U/ml Penicillin and 100 μg/ml Streptomycin) (Life Technologies, Darmstadt, Germany).

HEK 293T cells for transfections were purchased from Leibnitz Institute DSMZ (German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany) and cultivated at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium (Life Technologies, Darmstadt, Germany) supplemented with 10% fetal bovine serum (Biochrom, Berlin, Germany) and antibiotics (100 U/ml Penicillin and 100 μg/ml Streptomycin) (Life Technologies, Darmstadt, Germany). A total of 2.4×10^6 cells were seeded in 100-mm dishes and transiently transfected using PolyFect Transfection Reagent (Qiagen, Hilden Germany) according to the manufacturer's recommendations. In brief, 24 h after seeding, 8 μg of pSG5-miRNA expression plasmid DNA diluted in 300 μl of DMEM without supplements were used for the transfection. Cells and transfection complexes were incubated for 48 h at 37°C and 5% CO_2 to allow for miRNA overexpression.

RNA extraction

Total RNA including miRNA from cell lines was purified manually using miRNeasy Mini Kit (Qiagen, Hilden Germany) according to the manufacturer's protocol. Therefore, cell-culture DMEM was completely removed and the monolayer was carefully washed with 1 ml of phosphate buffered saline. About 700 μl of QIAzol Lysis Reagent was used to disrupt the cells by using a cell-scraper and vortexing. After adding 140 μl of chloroform, lysates were mixed thoroughly and centrifuged for 15 min at 12 000 *g* at 4°C to allow for phase separation. RNA was precipitated with a 1.5 vol. of 100% ethanol, washed and eluted in 2 × 40 μl H_2O RNase-free. Quality and quantity of isolated total RNA including miRNA were determined using NanoDrop 2000 UV-Vis Spectrophotometre (ThermoFisher Scientific, Waltham, Massachusetts, USA) with A260/280 ~2 and A260/230 ~ 1.8 and Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA) with RIN > 7.5.

Microarray analysis

miRNA abundance analysis of 12 samples was performed using Agilent microarrays for the Human miRBase V21 that contain probes for 2549 mature human miRNAs (Agilent Technologies). The procedures were performed as described previously according to the manufacturer's recommendations (34). A total of 100 ng total RNA from 12 cell lines (HEK 293T, PC-3, Tera-1, SHSY-5Y, HUH-7, DG-75, Jurkat, HeLa, JEG-3, MCF7, HaCaT and A549) (see also section 'Cell culture and transfection' for details) was processed using the miRNA Complete Labeling and Hyb Kit (Agilent Technologies) to generate fluorescently labeled miRNA. The microarrays were loaded and incubated at 55°C for 20 h with rotation. After washing, microarrays were scanned with the Agilent G2565CA Microarray Scanner System at 3 μm in double path mode. Raw data were acquired using Agilent AGW Feature Extraction software version 10.10.11 (Agilent Technologies). Background subtraction and quantile normalization of raw data were performed using R scripts (version 3.0.2) (35).

Northern blotting

For NB, 20 μg of total RNA including miRNA extracted by using miRNAeasy Mini Kit (Qiagen, Hilden Germany) was separated in 12% denaturing urea-polyacrylamide gels using SequaGel UreaGel System (National Diagnostics, Nottingham, UK) and 1x TBE running buffer. All buffers and solutions for NB were prepared using DEPC-treated H_2O to eliminate nuclease activity. A ssRNA marker was used to estimate the sizes of bands independently of the influence of external factors (temperature etc.) (RiboReady™ Color Micro RNA ladder, VWR, Radnor, PA, USA or Low range ssRNA Ladder and microRNA Marker, New England Biolabs, Frankfurt am Main, Germany). To check for loading control, the gel was stained with ethidiumbromide (10 mg/ml EtBr in 1× TBE) or 1× SYBR™ Gold in 1× TBE (Invitrogen/ThermoFisher Scientific, Waltham, Massachusetts, USA) and documented with a ChemiDoc Touch Imaging System (Bio-Rad, Munich, Germany). For semidry electroblotting, RNA was transferred to a Hybond N nylon membrane (GE Healthcare Life Sciences, Freiburg, Germany) for 30 min at 15 V. RNA was cross-linked chemically to the membrane using *N*-(3-Dimethylaminopropyl)-*N*′-ethylcarbodiimide hydrochloride (Sigma-Aldrich, Munich, Germany) for 2 h at 55°C. After cross-linking of endogenous RNA, the membranes were cut in half to prevent overlapping during hybridization. The generation of radiolabeled RNA probes was performed using miRVana miRNA Probe Construction Kit (Ambion/ThermoFisher Scientific, Waltham, Massachusetts, USA) following the manufacturer's instructions. Therefore, ssDNA templates composed of the full-length miRNA-of-interest sequence and an 8 nt T7 promoter sequence (5′-CCTGTCTC-3′) added to the 3′ end were hybridized to the T7 promoter primer. The remaining nucleotides corresponding to the template were added using Klenow DNA polymerase resulting in the desired dsDNA template. Next, the dsDNA template was *in vitro* transcribed using T7 RNA polymerase and radiolabeled UTP or GTP, if the probe was only containing two UTPs or less (Hartmann Analytic, Braunschweig, Germany). Template DNA was removed by DNase I digestion.

Pre-hybridization for 30 min was performed at 55°C in 5× SSC, 7% SDS, 1× blocking solution (Roche Diagnostics, Rotkreuz, Suisse), 20 mM Na_2HPO_4, 1× Denhardt's solution (Invitrogen/ThermoFisher Scientific, Waltham, Massachusetts, USA). Hybridization of radiolabeled probes to the cross-linked RNA membranes was performed at 55°C overnight. Blots were washed twice 15 min in 5× SSC, 1% SDS and twice 15 min in 1× SSC, 1% SDS at 55°C and exposed to a storage phosphor screen overnight. Screens were documented using a Typhoon scanner (GE Healthcare Life Sciences, Freiburg, Germany). Here, contrast and brightness were automatically adjusted by Typhoon Scanner Control Software (GE Healthcare Life Sciences, Freiburg, Germany) during scanning according to the darkest spot on the area. Whole NB images in this manuscript were further manually adjusted in terms of contrast and brightness.

Estimation of the validation rate

To provide an estimate of the total number of miRNAs, we considered the three groups, A: $p_1 = 897$ high-confidence miRNAs from the miRBase; B: $p_2 = 1759$ low-confidence miRNAs from the miRBase; C: $p_3 = 21\,471$ miRNAs not contained in the miRBase but predicted in various other studies separately from each other. For all three groups, we computed a high-throughput based exclusion rate based on the NGS data sets mentioned above (denoted as h_1, h_2, h_3, respectively) as well as a low throughput validation rate (l_1, l_2, l_3, respectively). We further assumed that high-throughput exclusion and low throughput validation are independent of each other. This makes the total estimated number of miRNAs to be: $\sum_{i=1}^{3} p_i \cdot (1 - h_i) \cdot l_i$.

RESULTS

Exclusion of likely false-positives by high-throughput data analysis

We collected 28 866 human small RNA sequencing samples containing 363.7 billion reads. Following stringent quality filtering, 8418 samples were excluded because of wrong annotations or low data quality. The remaining samples were mapped to 14 738 human miRNA precursor candidates totaling 24 094 mature miRNA candidates. These can be split in three basic sets: A: $p_1 = 897$ high-confidence miRNAs from miRBase V22; B: $p_2 = 1759$ low-confidence miRNAs from miRBase V22; C: $p_3 = 21\,471$ miRNAs not contained in miRBase but predicted in various other studies. For mapping of those billions of small RNA sequencing reads, only 4.9% of high-confident miRNAs did not fulfill the quality criteria ($h_1 = 0.049$). For low-confidence miRNAs from miRBase, this rate already increased to 38.6% (h_2

= 0.386). For the set of miRNA candidates, not yet annotated in miRBase, this rate increased to 70.1% ($h_3 = 0.701$). In total, 853 of 897 high-confidence miRBase miRNAs, 1080 of 1759 low-confidence miRBase miRNAs and 6412 of 21 471 novel candidates fulfilled the high-throughput criteria (details are provided in Supplementary Table S1).

Detection of IsomiR variants

For 2873 miRNAs from miRBase V22, we found at least one isomiR variant with additional nucleotides either at their 5p or at their 3p end (Supplementary Table S3). From our training set, miR-6829–5p was found to have the highest number (16) of isomiRs detected. miR-34a-3p showed the most isomiRs (9) from our positively validated miRNA set. The sequences and corresponding reads per million mapped (RPMMM) are listed in Table 2. The standard sequence for miR-34a-3p, designated as '0F_0T' (zero changes at the five prime end and zero changes at the three prime end) was represented by 34.75% RPMMM. The next highest number of RPMMM (16.52%) was found for variant '0F_1T' with one additional nucleotide at the 3p end. The third highest number of RPMMM (13.79%) was found for isomiR variant '1F_1T' with one additional nucleotide at the 5p and one additional nucleotide at the 3p end. MiRNA isomiR variants were not further validated by northern blotting due to the lack of specificity of RNA probes for the nucleotide changes at 5p or 3p ends.

Validation of exogenous miRNAs by northern blotting

To select miRNAs for experimental validation, we employed a recently developed algorithm that acknowledges key criteria characteristic for miRNAs. The complete listing of criteria to define high-confident miRNAs is given in Backes *et al.* (20) (see also our implemented web-server for ranking potential miRNA candidates at www.ccb.uni-saarland.de/novomirank). After excluding precursors that did not meet the criteria of high-confident miRNAs, we employed an experimental validation step to further identify false-positive miRNAs. From each of the three abovementioned groups, we selected a representative number for low-throughput validation as described in the 'Materials and Methods' section. Altogether, 203 mature miRNAs (originating from 108 precursor molecules) were tested in this validation step. To this end, miRNA precursor sequences were cloned into pSG5-miRNA expression plasmids and recombinantly expressed in HEK 293T cells. Since the likelihood for true positives was highest in set A and lowest in set C, the test set sizes were selected accordingly, with respect to miRBase V21, which was the most recent version at the time of the study setup: we analyzed 51 (40 in V21) high-confidence miRNAs from miRBase V22 (set A), 33 (41 in V21) low-confidence miRNAs from miRBase V22 (set B) and 119 (122 in V21) novel miRNA candidates from set C. Out of the three miRNA candidates that were not present in V21 yet, one (hsa-miR-9903) has been added to the high-confidence set and was successfully validated by us, while the other two (hsa-miR-12129, hsa-miR-10527–5p) were added to the low-confidence set and did not pass our validation. Notably, the three miRNA precursor candidates predicted by us to encode both 5p and 3p miRNA have been added to miRBase V22 as giving rise to only one mature form. In addition to miR-9903–3p, miR-9903–5p was also positively validated by us and should be included in miRBase.

Validation of endogenous miRNAs by northern blotting

To search for endogenous miRNAs with expression levels high enough to be identifiable by northern blotting, we analyzed 12 human cell lines by microarray. In detail, miRNA abundance analysis was performed using Agilent microarrays based on miRBase V21 that contains probes for 2549 mature human miRNAs. The 12 cell lines were derived from prostate (PC-3), testis (Tera-1) bone marrow (SHSY-5Y), liver (HUH-7), lung (A549), B (DG-75)- and T-lymphocytes (Jurkat), cervix (HeLa), placenta (JEG-3), mammary gland (MCF7), keratinocytes (HaCaT) and HEK 293T (kidney) as a reference. Based on the array results, we selected 11 miRNAs with high expression levels to be detectable by northern blotting (see Table 3 for array data of selected miRNAs and Supplementary Table S3 for array data of V22 miRNAs). Notably, the miRNAs, miR-4284 and miR-1260a that were upon the most recently deposited miRNAs, i.e. with a very high ID, did not yield signals of the expected size in any of the analyzed cell lines. miR-7975 that, according to the microarray, showed the strongest expression in all 12 cell lines exhibited several potential signals but no precise assignment to mature or precursor forms was possible. All other miRNAs including 137–3p, 148a-3p, 155–5p, 16–5p, 19b-3p, 20a-5p, 23a-3p and 23b-3p yielded signals in the expected size range for the mature form as shown in Figure 2. While the miRNAs 16–5p, 19b-3p, 20a-5p, 23a-3p and 23b-3p were detected in all 12 cell lines, miR-137–3p was found in 5 cell lines only, miR-148a-3p in 10 cell lines and miR-155–5p in 5 cell lines (Figure 2B). Out of the 11 miRNAs, the miRNAs 137–3p, 148a-3p, 155–5p, 23a-3p and 23b-3p were also identified by the exogenous expression analysis.

Taken together, from our exogenous and endogenous experiments, we successfully validated 51 of 54 high-confidence miRNAs (94.4%; $l_1 = 0.944$), 10 of 35 low-confidence miRNAs (28.6%, $l_2 = 0.286$) and 22 of 119 miRNAs that have not yet been annotated in miRBase V22 (18.5%, $l_3 = 0.185$). Details are provided in Supplementary Table S2. Our simplified model described in the 'Materials and Methods' section lets us estimate that 806 of the 897 miRBase high-confidence miRNAs are true positives (89.8%), 309 of the 1759 low-confidence miRNAs (17.5%) and 1185 of the 21 471 potential candidates (5.5%).

Specific analysis of mature and precursor miRBase miRNAs

Out of the 89 miRNAs from miRBase V22 (54 high- and 35 low-confidence), we detected northern blot signals for 61 mature miRNAs (51 high- and 10 low-confidence). All of these signals for mature miRNAs approximately matched the expected size range. Most of them were not discovered in the endogenous controls. mir-1260a and mir-23c showed two signals in the size range of their mature miRNAs, signal intensities virtually corresponded to those found in control cells. We defined miRNAs as positive when signals for

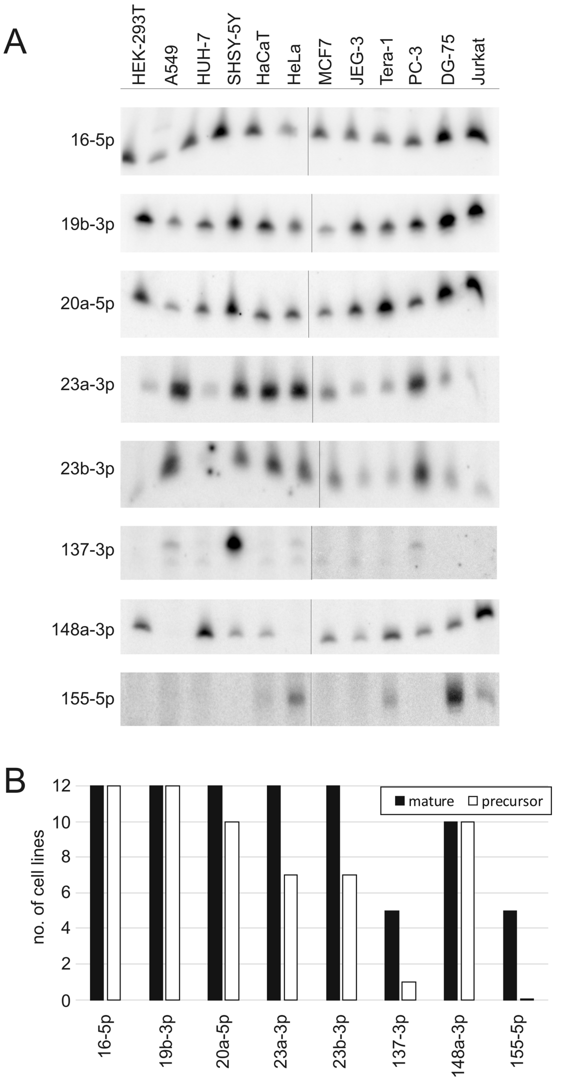


Figure 2. Northern blots of endogenous miRNAs. (**A**) The 11 analyzed cell lines that are indicated on top of the figure were derived from lung (A549), liver (HUH-7), bone marrow (SHSY-5Y), keratinocytes (HaCaT), cervix (HeLa), mammary gland (MCF7), placenta (JEG-3), Testis (Tera-1), prostate (PC-3), B-lymphocytes (DG-75) and T-lymphocytes (Jurkat). HEK-293T RNA was used as a reference to compare signals to exogenously expressed miRNAs. The endogenous mature forms are shown for the miRNAs indicated on the left side of the figure. (**B**) The number of mature and premature forms of the endogenous miRNA expressed in the 12 cells lines as indicated in Figure 2A.

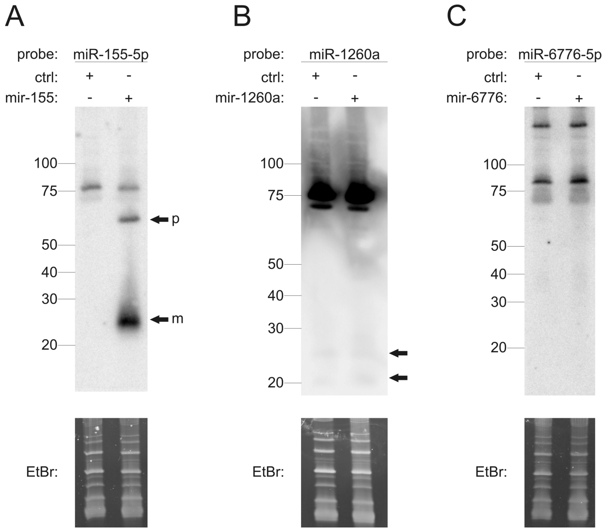


Figure 3. Representative NB results for a positive, questionable and negative miRNA in HEK 293T cells. (**A**) NB for hsa-miR-155–5p from high-confidence set A showing distinct bands for its precursor (p) and mature (m) form. (**B**) Hybridization against hsa-miR-1260a (low-confidence set B) detects two small RNA fragments with similar signal intensities for the control. (**C**) Probing for hsa-miR-6776–5p (low-confidence set B) did not result in any specific bands. Ethidium bromide staining of RNA gels was used as a loading control.

both a precursor and a mature form were detected and are stronger for the overexpression compared to control RNA lysates. However, in the majority of cases, the size of the precursor did not correspond to the size indicated for the respective stem-loop forms by miRBase. NB examples for positive, negative and doubtful miRNAs from miRBase V22 are shown in Figure 3.

Considering all cases, we found six recombinants and two endogenous miRNAs, as already described above, with signals neither in the size range of the precursor nor in the range of the mature forms. The only high-confident miRNAs that were not confirmed by our northern blot analysis were miR-6511a-5p and miR-6511a-3p, both of which showed signals for potential precursor forms, albeit in different sizes, but not for any of both mature forms and miR-26b-3p where only a premature form could be detected. These results suggest that even in the high-confidence set a certain number of false-positive miRNAs exist.

For the low-confident miRNAs, exogenous and endogenous expression analysis failed to confirm 22 and 3 miRNAs, respectively, that did not show signals for the mature and precursor form. About 19 of these had at least some signals designated as potential precursor form or unclear signals (details provided in Table 1). The remaining 10 miRNAs were confirmed by the identification of both the mature and the precursor forms. Although our analysis is largely consistent with the miRBase qualification of many low-confident miRNAs, our data also indicate a considerable number of false-negative miRNAs among the low-confident set in miRBase.

Ratio of precursor and mature -5p/-3p miRNA forms

As for variances in signal intensities, 49 miRNAs showed stronger signals for the precursor than for the mature form indicating a reduced processing efficiency (Table 1). As abovementioned, miR-6511a-5p and 24 other miRNAs showed only signals for the precursor but not for the mature form. Overall, the 5p-forms appear to be more efficiently processed into mature miRNAs than the 3p-forms. Out of the 14 miRNAs, which gave rise to one mature form only

Table 1. Comparison between NB signal intensities of 5p versus 3p mature miRNA and corresponding precursor forms from set A (high-confidence), set B (low-confidence) and set C (miRNA candidates)

set	miRNA (candidate)	pre	mat	pre versus mat	pre	mat	pre versus mat
A	10a	moderate	strong	weaker	weak	strong	weaker
A	(endo) 16	moderate	strong	weaker	–	–	–
A	(endo) 19b	–	–	–	weak	strong	weaker
A	(endo) 20a	–	–	–	weak	strong	weaker
A	23a	moderate	weak	stronger	weak	strong	weaker
A	26b	moderate	moderate	weaker	moderate	n. d.	pre only
A	27b	weak	weak	equal	strong	weak	stronger
A	34a	moderate	strong	weaker	moderate	weak	stronger
A	101	weak	near bg	stronger	moderate	weak	stronger
A	122	weak	strong	weaker	moderate	moderate	stronger
A	125a	weak	strong	weaker	moderate	strong	weaker
A	137	–	–	–	moderate	strong	weaker
A	140	weak	weak	weaker	weak	weak	weaker
A	142	moderate	moderate	stronger	near bg	moderate	weaker
A	143	moderate	moderate	equal	strong	weak	stronger
A	145	weak	strong	weaker	weak	weak	weaker
A	148a	strong	weak	stronger	weak	strong	weaker
A	(endo) 148a	–	–	–	weak	moderate	weaker
A	155	moderate	strong	weaker	moderate	weak	stronger
A	181a	moderate	strong	weaker	weak	weak	stronger
A	191	weak	near bg	stronger	strong	near bg	stronger
A	193a	moderate	moderate	stronger	weak	strong	weaker
A	195	moderate	strong	weaker	weak	weak	stronger
A	205	strong	moderate	stronger	strong	strong	equal
A	301a	weak	weak	stronger	moderate	strong	weaker
A	361	moderate	strong	weaker	moderate	weak	stronger
A	375	–	–	–	moderate	strong	weaker
A	483	moderate	moderate	equal	strong	strong	stronger
A	497	weak	strong	weaker	weak	near bg	stronger
A	874	weak	weak	stronger	near bg	near bg	equal
A	6511a	moderate	n. d.	pre only	weak	n. d.	pre only
A	9903	strong	strong	weaker	weak	strong	weaker
B	23b	moderate	weak	stronger	weak	strong	weaker
B	#23c	–	–	–	strong	weak	stronger
B	#133b	–	–	–	moderate	strong	weaker
B	#630	–	–	–	near bg	n. d.	pre only
B	#665	–	–	–	near bg	n. d.	pre only
B	939	n. d.	n. d.	–	n. d.	n. d.	–
B	#1202	weak	near bg	stronger	–	–	–
B	1228	weak	n. d.	pre only	weak	n. d.	pre only
B	1229	moderate	n. d.	pre only	strong	weak	stronger
B	1238	weak	weak	equal	weak	n. d.	pre only
B	#1246	weak	near bg	stronger	–	–	–
B	#1260a	n. d.	near bg	mat only	–	–	–
B	#3137	weak	weak	equal	–	–	–
B	#3148	moderate	strong	weaker	–	–	–
B	3162	weak	moderate	weaker	n. d.	n. d.	–
B	#4534	–	–	–	weak	n. d.	pre only
B	#4721	–	–	–	moderate	n. d.	pre only
B	6776	n. d.	n. d.	–	n. d.	near bg	mat only
B	6829	near bg	n. d.	pre only	weak	n. d.	pre only
B	6865	n. d.	moderate	mat only	weak	n. d.	pre only
B	10527	weak	moderate	weaker	n. d.	moderate	mat only
B	12129	n. d.	n. d.	–	n. d.	n. d.	–
C	novel-241	n. d.	n. d.	–	n. d.	n. d.	–
C	novel-1002	n. d.	n. d.	–	n. d.	n. d.	–
C	novel-1037	n. d.	n. d.	–	n. d.	n. d.	–
C	novel-1043	near bg	n. d.	pre only	weak	near bg	weaker
C	novel-1236	weak	moderate	stronger	weak	n. d.	pre only
C	novel-1521	n. d.	n. d.	–	near bg	n. d.	pre only
C	novel-1564	n. d.	moderate	mat only	n. d.	near bg	mat only
C	novel-1790	weak	near bg	stronger	n. d.	n. d.	–
C	novel-1887	near bg	weak	weaker	weak	n. d.	pre only
C	novel-2295	moderate	weak	stronger	weak	weak	weaker
C	pnm-18	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-339	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-501	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-1089	near bg	strong	weaker	near bg	strong	weaker
C	pnm-1523	n. d.	weak	mat only	n. d.	near bg	mat only

Table 1. Continued

set	miRNA (candidate)	pre	mat	pre versus mat	pre	mat	pre versus mat
C	pnm-1609	weak	n. d.	pre only	n. d.	n. d.	–
C	pnm-1728	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-2012	moderate	near bg	stronger	weak	moderate	weaker
C	pnm-2523	near bg	near bg	stronger	near bg	near bg	stronger
C	pnm-2908	n. d.	near bg	mat only	n. d.	n. d.	–
C	pnm-3375	n. d.	weak	mat only	n. d.	n. d.	–
C	pnm-4607	moderate	strong	weaker	n. d.	near bg	mat only
C	pnm-4828	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-4927	weak	near bg	stronger	strong	moderate	stronger
C	pnm-6141	n. d.	n. d.	–	n. d.	near bg	mat only
C	pnm-6147	n. d.	n. d.	–	weak	near bg	stronger
C	pnm-6785	n. d.	n. d.	–	weak	near bg	stronger
C	pnm-7379	weak	moderate	weaker	moderate	moderate	equal
C	pnm-7519	moderate	weak	stronger	n. d.	near bg	mat only
C	pnm-8500	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-8679	strong	moderate	stronger	weak	weak	equal
C	pnm-8692	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-8893	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-9262	weak	near bg	stronger	n. d.	n. d.	–
C	pnm-10387	n. d.	n. d.	–	near bg	near bg	equal
C	pnm-10468/-71/-72	n. d.	n. d.	–	n. d.	near bg	mat only
C	pnm-10470	n. d.	n. d.	–	n. d.	near bg	mat only
C	pnm-10565	strong	near bg	stronger	n. d.	near bg	mat only
C	pnm-10945	n. d.	n. d.	–	near bg	near bg	stronger
C	pnm-11436	weak	near bg	stronger	weak	near bg	stronger
C	pnm-11712	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-12346	n. d.	near bg	mat only	n. d.	n. d.	–
C	pnm-12352	weak	near bg	stronger	weak	near bg	stronger
C	pnm-12395	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-13945	weak	weak	equal	moderate	moderate	equal
C	pnm-14137	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-14397	n. d.	n. d.	–	n. d.	near bg	mat only
C	pnm-15272	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-15546	n. d.	weak	mat only	n. d.	weak	mat only
C	pnm-16556	strong	near bg	stronger	moderate	near bg	pre only
C	pnm-17724	weak	moderate	weaker	strong	weak	stronger
C	pnm-20077	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-20714	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-21981	near bg	n. d.	pre only	n. d.	near bg	mat only
C	pnm-22155	n. d.	n. d.	–	n. d.	n. d.	–
C	pnm-22472	weak	n. d.	pre only	weak	n. d.	pre only
C	pnm-23093	strong	n. d.	pre only	strong	near bg	stronger
C	pnm-23453	weak	near bg	stronger	strong	n. d.	pre only

#miRNAs are predicted to only give rise to one mature form. n. d.: not detectable, near bg: near background. miRCarta IDs of novel and miRA candidates were shown in Supplementary Table S5.

according to miRBase V22, we confirmed 4 miRNAs, 2 of which, including mir-133b and mir-3148, showing stronger signals for their mature form than for the precursor (Table 1).

Failed confirmation of miRNAs in recent miRBase releases

Grouping miRNAs to the according miRBase version of which the miRNA has first been listed highlights that mature miRNAs confirmed by our NB based approach are enriched in early miRBase versions. In detail, all 33 miRNAs that were taken from miRBase versions 1 through 7 have been experimentally validated. Version 8 of miRBase was the first version to contain a miRNA that showed no clear signals. Here, for mir-630, we only detected a very faint signal in precursor size. Out of 26 miRNAs taken from version 10, 20 have been verified by our NB analysis. Only 3 out of 20 miRNAs taken form version 17 onwards have been verified. Consistent with the doubts raised in previous studies, these results further question the nature of miRNAs with high ID numbers recently deposited in miRBase that are mostly identified by NGS studies only.

Stability of estimated human miRNome size between miRBase V21 and V22

At the time of the study setup miRBase V22 was not yet released. Therefore, we selected miRNAs for analysis in accordance to the proportions of the high- and low-confidence miRNA annotations of miRBase V21, containing 544 and 2044 miRNAs respectively. Extrapolating the miRNome size based on the data of miRBase V21, 509 miRNAs of the high-confidence set were considered true positives (93.6%), 627 of the low-confidence miRNAs (30.7%) and 1213 of

Table 2. Relative isomiR read counts (> 2%) detected for hsa-miR-34a-3p

		Relative RPMMM		
sequence	iso_type	(%)	Total RPMMM	Total reads
CAAUCAGCAAGUAUACUGCCC	0F_-1T	4.00	2079.53	13 287
CAAUCAGCAAGUAUACUGCC	0F_-2T	2.10	1089.15	6082
CAAUCAGCAAGUAUACUGCCCU	0F_0T	34.75	18 060.38	93 848
CAAUCAGCAAGUAUACUGCCCUA	0F_1T	16.52	8585.07	46 920
CAAUCAGCAAGUAUACUGCCCUAG	0F_2T	6.09	3164.45	19 520
AAUCAGCAAGUAUACUGCCCU	1F_0T	8.90	4623.40	22 909
AAUCAGCAAGUAUACUGCCCUA	1F_1T	13.79	7167.57	40 377
AAUCAGCAAGUAUACUGCCCUAG	1F_2T	5.87	3053.31	21 018
AUCAGCAAGUAUACUGCCCUAG	2F_2T	3.20	1664.71	9701

Iso-types are defined as additional or absent nucleotides on 5′ (F) or 3′ (T) ends, respectively. RPMMM = Reads Per Million Mapped to MiRNAs. Iso-type 0F_0T represents the commonly known sequence from miRBase V22.

Table 3. Quantile normalized microarray data for selected miRNAs in 12 cell lines

	HEK			SHSY-								
miRNA	293T	A549	HUH7	5Y	HaCaT	HeLa	MCF7	JEG3	Tera1	PC3	DG75	Jurkat
hsa-miR-16–5p	9.733	9.428	9.733	9.946	9.501	9.428	9.946	7.474	8.832	9.213	10.403	11.161
hsa-miR-19b-3p	11.404	9.036	10.220	10.756	9.946	10.220	8.729	9.733	10.220	9.309	11.161	11.404
hsa-miR-20a-5p	10.936	8.832	10.062	10.550	9.863	9.501	8.609	9.428	9.863	9.036	10.936	10.936
hsa-miR-23a-3p	4.952	9.733	7.036	8.454	9.733	10.062	8.522	5.197	7.182	9.428	7.595	5.986
hsa-miR-23b-3p	5.578	9.213	6.527	7.893	7.437	8.177	7.437	4.101	4.481	7.256	5.777	5.219
hsa-miR-137–3p	0.583	4.467	2.095	7.474	1.728	4.345	-0.094	0.148	0.444	2.142	0.208	0.118
hsa-miR-148a-3p	7.523	1.577	8.123	5.049	4.710	1.636	6.445	3.961	7.394	6.384	6.527	8.307
hsa-miR-155–5p	-0.300	0.246	-0.062	-0.143	1.118	3.984	-0.235	0.256	3.346	0.263	5.435	3.612
hsa-miR-1260a	11.891	11.161	11.891	11.891	11.404	11.404	11.891	11.891	10.936	11.404	11.404	10.756
hsa-miR-4284	11.161	10.936	11.404	11.161	10.062	11.161	11.161	9.636	10.062	10.550	9.863	10.403
hsa-miR-7975	16.004	16.004	16.004	16.004	16.004	16.004	16.004	16.004	16.004	16.004	16.004	16.004

the candidate miRNAs (5.6%), resulting in an estimated miRNome size of 2349 miRNAs. The release of miRBase V22 had a large impact on the proportions of the high- and low-confidence sets, increasing the size of the high-confidence set by 65% to 897 miRNAs. While the proportions changed substantially, our estimate remained stable and decreased only by 49 miRNAs (2.1%).

Likely false-positive miRNAs and their targets

There were no miRNA targets for miRNAs that were predicted in various studies but not deposited in miRBase (set C). By contrast, target genes have been annotated for almost all miRNAs taken from miRBase V21. At the time of writing, no annotations were available for miRNAs from miRBase V22. Here, we compared the targetomes of miRNAs verified by our above applied criteria versus the targetomes of miRNAs not confirmed in our analysis. To this end, we considered only miRNA targets that have been classified as verified targets by miRTarBase (strong evidence). In the set of validated miRNAs, each miRNA had a median of 144 validated target genes. In the set of the not-validated miRNAs, each had a decreased number of 65 (median) target genes. While the difference was statistically significant (P=0.009) as determined by Wilcoxon Rank-sum test, these results demonstrate substantial numbers of targets for not-validated miRNAs. For example, both mature forms of hsa-mir-939 that we could not validate in this study have been associated with complex target gene sets of 199 for the 5p and 439 targets for the 3p form. Altogether, 16 miRNAs not-validated by our approach have recently been associated with 2844 experimentally validated targets.

DISCUSSION

Messenger RNA transcripts have been frequently confirmed as endogenous molecules by NB. Due to their rather low endogenous expression few endogenous miRNAs are identifiable by NB. Although there are no studies that systematically analyzed endogenous miRNAs by NB, there are data sets listing endogenous miRNAs according to their signal intensities (e.g. GSM1513689 deposited in GEO). The overall low endogenous expression of miRNAs necessitates an exogenous expression system, which also allows monitoring the processing of the precursor into the mature form. We chose HEK 293T cell culture as expression system that stems from a kidney of a healthy aborted fetus and that allows high transfection efficiency and high expression rates for the pSG5 vector (36,37). The identification of both the precursor and the mature form by NB indicates processing of a miRNA and strongly argues in favor of a true miRNA.

The use of an exogenous expression system also allowed to systematically compare the processing of the 5p and the 3p form for each of the analyzed miRNAs. Depending on the tissue, the cell, and the applied condition, both mature forms have been reported as functional (38). In case of miRNAs for which no mature but a precursor form can be detected by our NB procedure, it is conceivable that the amount of miRNA was too low to yield distinct signals. Overall, our data indicated that the 5p-forms are more frequently processed into mature miRNAs than the 3p-forms. This observation is consistent with previous publications that analyzed miRNA strand selection with regards to thermodynamic stability (39,40). Although our validation pipeline was not optimized for the detection of splicing-derived miRNAs, we obtained positive NB signals for mirtron precursor mir-1229 and its mature 3p form but not for mir-1228 and mir-1238. To the best of our knowledge, the only studies that also identified mirtrons by NB were by Schamberger *et al.* (41), who reported mirtron mir-1226 as one out of three candidates tested and Hubé *et al.* (42) who detected 6 out of 56 short intron derived miRNA candidates. Agotrons, another potential exception for our validation system, do not appear as miRNA like signals or even not as a premature like form on northern blots as they differ in size up to 100 nt and irrespective of their association with Ago proteins. For the specific validation of exogenous agotrons via northern blotting they should be co-expressed along with Argonaute proteins for stabilizing effects (43). In contrast to the detection of single-nucleotide polymorphisms, isomiR detection is not readily possible by quantitative Real Time-PCR or northern blotting. These methods do not allow to discriminate between miRNAs with length variants of up to 5 nt at their 5p or 3p ends. Almost all of the studies describing isomiR variants use enzyme-based methods making the bias-free validation of miRNA isoforms that were detected by NGS a major challenge (44–46).

For a substantial number of miRNAs, we failed to establish confirmation by NB. This should raise the awareness that a presumed miRNA may not be a true miRNA, even if other methods, i.e. microarray or qRT-PCR, suggest high expression values or other studies already provided evidence for its functionality. For example, we failed to validate mir-939 exogenously and miR-4284 and miR-7975 endogenously, although others have already provided functional evidence for its derived miRNAs (47–53 and many others). Failed confirmation by NB does however not necessarily disqualify a miRNA as true miRNA. As addressed above, one has to differentiate between miRNAs that show a signal for the precursor miRNA only but no processing, a signal only for the processed form, and cases without any or with only faint signals. While the latter cases likely do not represent true miRNAs, even the lack of identification of the two forms does not necessarily disprove that a tested sequence represents a true miRNA. Tissue specific factors that are required for processing in a given cell may not be present in the HEK cells used (21). Finally, signal intensities also depend on the amount of miRNA expressed and processed and the number of radiolabeled nucleotides in probes. To minimize the influence of biased labeling, we used radiolabeled GTP for all oligonucleotide probes that contained only two or less UTPs.

The abovementioned limitations are also found in other systems like knockout systems for Drosha/Dicer that have been used to confirm true canonical miRNAs. In this system, a failure of processing a miRNA precursor in the Drosha/Dicer mutants may be due to the specific physiology of these mutants. A failure of processing a miRNA precursor in the according wild-type cells can likewise be linked to a specific cell type. In addition, most of the studies that use knockouts for Drosha/Dicer characterize the analyzed miRNA candidates by NGS and PCR entailing the problems related to these techniques. In sum, knockout experiments of proteins processing miRNAs do not rule out erroneous confirmation of false-positive miRNAs or erroneous disproval of real miRNAs (54).

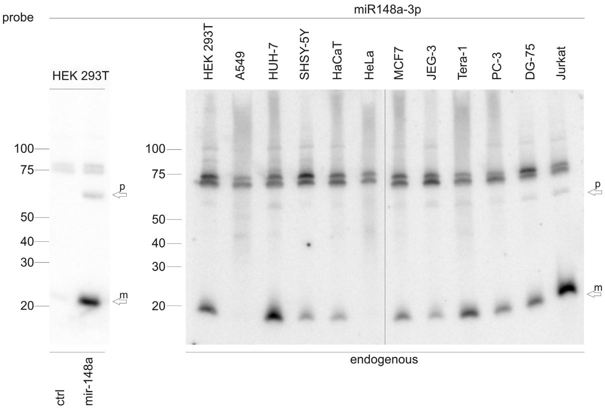


Figure 4. Comparison of exogenous and endogenous miRNA expression by northern blotting**.** The mature form is indicated by 'm' and the premature form by 'p'. The left part of the figure shows exogenous expression of miR-148a-3p in HEK 293T cells. HEK 293T cells that were transfected with an empty vector are shown as a control (ctrl). The right part of the figure shows endogenous expression of miR-148a-3p in 12 cells lines as specified in Figure 2A. To show endogenous miRNAs, the signal intensity of the endogenous miRNAs apparently is enhanced as compared to the signal intensity of the exogenous miRNAs due to the very strong signal of overexpressed miR-148a-3p (compare backgrounds of exo- and endogenous northern blots and see the comparison between HEK 293T cells that were used as a control for the transfection analysis (ctrl) shown in the left part of the figure and the HEK 293T cells that were compared with other cells shown in the right part of the figure). As described in the 'Materials and Methods' section, contrast and brightness were adjusted by the software during scanning according to the darkest spot on the blot.

Complementary to experimental settings that manipulate miRNA expression of a specific cell type, endogenous miRNA expression can be analyzed. To this end, we tested 11 human cell lines in addition to HEK 293T cells by northern blotting and found signals for several miRNAs in the size range of the mature sequences. For endogenously detected precursors, the sizes varied to a degree comparable to the range detected for the exogenous analyses. Notably, and as shown in Figure 4, the signal sizes of the endogenous miRNAs (mature and premature forms) correspond to those of the exogenous miRNAs further supporting the validity of the data obtained for the induced miRNAs. The relatively weak intensity of the signals of the endogenous miRNAs show, however, that only a minor portion of the miRNAs can be identified without induced overexpression.

Most miRNAs that have failed the northern blot validation step were detected by less NGS reads and in less NGS samples compared to miRNAs that passed this step. However, there are also examples of miRNAs that were highly expressed in many NGS samples or microarray data sets but still failed validation by northern blotting. For example, miR-26b-3p, miR-6511a-3p and miR-1246 were all detected in over 10 000 samples, with in total more than 230 000 reads per miRNA. In addition, three miRNAs, i.e. miR-7975, miR-1260a and miR-4284, which were upon the top 16 miRNAs expressed in all 12 cell lines analyzed, failed the northern blot filtering. Furthermore, the opposite was also found. Among the validated miRNAs, miR-1202, miR-3148, miR-3137 and miR-3162–5p were all detected in <100 samples, when requiring at least 10 reads per sample. Accordingly, a read number based filtering would surely not be sufficient to eliminate false-positive miRNAs.

In the light of the different possibilities to identify and confirm miRNAs, it is certainly not justified to prematurely limit a quality control scheme to few methods. NB is certainly only one quality criterion for a true miRNA but nevertheless an essential one. Arguments for NB as a method to confirm true miRNAs are the possibility of a high degree of standardization, the visual demonstration of the product sizes, the omission of amplification and ligations steps, and the possibility to show processing from the precursor into the mature form by using an exogenous expression system. We strongly suggest including NB in future database formats. To provide highest data integrity, data repositories with NB data should provide (i) complete documentation of the NB by scanned images with brightness/contrast settings without image adjustments, (ii) the method of RNA extraction, (iii) the size of the primary transcript that is (over-) expressed in the cell type analyzed, (iv) the minimal sequence surrounding a miRNA hairpin that assures correct processing of mature miRNAs, (v) the respective cells used for validation analyses and (vi) if applicable endogenous NB signal intensities and their correlation with qRT-PCR data and/or microarray and/or NGS data.

In summary, we found the highest number of confirmed miRNAs for the high-confidence set of miRNAs of miRBase V22, particularly among the miRNAs deposited in miRBase releases 1 through 10. We also present a set of miRNAs with a precursor and the mature forms confirmed by the NB pipeline, which have not yet been annotated in miRBase. In total, our data indicate 2300 human mature miRNAs ~50% (1115) of which are annotated in miRBase V22. Since the high- and low-throughput validation presented in this study are not independent of each other, our model represents certainly only an estimate of the total number of miRNAs. Based on the presented data, our estimation, however, likely indicates the upper number of what we can expect as the final extent of the true human miRNome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: J.A., U.F. and M.M. performed NB experiments. J.A. performed miRNA microarrays and edited raw NB images. M.H. and M.A.-H. prepared expression plasmids for 22 and 6 miRNAs, respectively. F.A.G. provided expression plasmids for 26 miRNAs and 4 cell lines. T.F., C.B. and V.G. performed the bioinformatics analyses. V.G. prepared the graphical abstract and Figure 1. J.A., T.F., H.-P.L., A.K. and E.M. wrote the manuscript.

FUNDING

Michael J. Fox Foundation (US), Grant ID: 14446. Funding for open access charge: Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
3. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
4. Hart,M., Nolte,E., Wach,S., Szczyrba,J., Taubert,H., Rau,T.T., Hartmann,A., Grasser,F.A. and Wullich,B. (2014) Comparative microRNA profiling of prostate carcinomas with increasing tumor stage by deep sequencing. *Mol. Cancer Res.*, **12**, 250–263.
5. Petriella,D., De Summa,S., Lacalamita,R., Galetta,D., Catino,A., Logroscino,A.F., Palumbo,O., Carella,M., Zito,F.A., Simone,G. *et al.* (2016) miRNA profiling in serum and tissue samples to assess noninvasive biomarkers for NSCLC clinical outcome. *Tumour Biol.*, **37**, 5503–5513.
6. Drusco,A., Nuovo,G.J., Zanesi,N., Di Leva,G., Pichiorri,F., Volinia,S., Antenucci,A., Costinean,S., Bottoni,A. *et al.* (2014) MicroRNA profiles discriminate among colon cancer metastasis. *PLoS One*, **9**, e96670.
7. Alles,J., Menegatti,J., Motsch,N., Hart,M., Eichner,N., Reinhardt,R., Meister,G. and Grasser,F.A. (2016) miRNA expression profiling of Epstein-Barr virus-associated NKTL cell lines by Illumina deep sequencing. *FEBS Open Bio.*, **6**, 251–263.
8. Wen,Y., Han,J., Chen,J., Dong,J., Xia,Y., Liu,J., Jiang,Y., Dai,J., Lu,J., Jin,G. *et al.* (2015) Plasma miRNAs as early biomarkers for detecting hepatocellular carcinoma. *Int. J. Cancer*, **137**, 1679–1690.
9. Akers,J.C., Hua,W., Li,H., Ramakrishnan,V., Yang,Z., Quan,K., Zhu,W., Li,J., Figueroa,J., Hirshman,B.R. *et al.* (2017) A cerebrospinal fluid microRNA signature as biomarker for glioblastoma. *Oncotarget*, **8**, 68769–68779.
10. Abu-Halima,M., Meese,E., Keller,A., Abdul-Khaliq,H. and Radle-Hurst,T. (2017) Analysis of circulating microRNAs in patients with repaired Tetralogy of Fallot with and without heart failure. *J. Transl. Med.*, **15**, 156.
11. Keller,A. and Meese,E. (2016) Can circulating miRNAs live up to the promise of being minimal invasive biomarkers in clinical settings? *Wiley Interdiscip. Rev. RNA*, **7**, 148–156.
12. Ludwig,N., Nourkami-Tutdibi,N., Backes,C., Lenhof,H.P., Graf,N., Keller,A. and Meese,E. (2015) Circulating serum miRNAs as potential biomarkers for nephroblastoma. *Pediatr. Blood Cancer*, **62**, 1360–1367.
13. Abu-Halima,M., Hammadeh,M., Backes,C., Fischer,U., Leidinger,P., Lubbad,A.M., Keller,A. and Meese,E. (2014) Panel of five microRNAs as potential biomarkers for the diagnosis and assessment of male infertility. *Fertil. Steril.*, **102**, 989–997.
14. Leidinger,P., Backes,C., Deutscher,S., Schmitt,K., Mueller,S.C., Frese,K., Haas,J., Ruprecht,K., Paul,F., Stahler,C. *et al.* (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.*, **14**, R78.
15. Keller,A., Leidinger,P., Bauer,A., Elsharawy,A., Haas,J., Backes,C., Wendschlag,A., Giese,N., Tjaden,C., Ott,K. *et al.* (2011) Toward the blood-borne miRNome of human diseases. *Nat. Methods*, **8**, 841–843.
16. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
17. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111.
18. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
19. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
20. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grasser,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
21. Chiang,H.R., Schoenfeld,L.W., Ruby,J.G., Auyeung,V.C., Spies,N., Baek,D., Johnston,W.K., Russ,C., Luo,S., Babiarz,J.E. *et al.* (2010)

Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.*, **24**, 992–1009.

22. Meng,Y., Shao,C., Wang,H. and Chen,M. (2012) Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants. *RNA Biol.*, **9**, 249–253.
23. Wang,X. and Liu,X.S. (2011) Systematic curation of miRBase annotation using integrated small RNA High-Throughput sequencing data for C. elegans and drosophila. *Front. Genet.*, **2**, 25.
24. Hansen,T.B., Kjems,J. and Bramsen,J.B. (2011) Enhancing miRNA annotation confidence in miRBase by continuous cross dataset analysis. *RNA Biol.*, **8**, 378–383.
25. Brown,M., Suryawanshi,H., Hafner,M., Farazi,T.A. and Tuschl,T. (2013) Mammalian miRNA curation through next-generation sequencing. *Front. Genet.*, **4**, 145.
26. Fromm,B., Billipp,T., Peck,L.E., Johansen,M., Tarver,J.E., King,B.L., Newcomb,J.M., Sempere,L.F., Flatmark,K., Hovig,E. *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.*, **49**, 213–242.
27. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.P., Meese,E. and Keller,A. (2017) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
28. Krawczak,M., Reiss,J., Schmidtke,J. and Rosler,U. (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucleic Acids Res.*, **17**, 2197–2201.
29. Meyerhans,A., Vartanian,J.P. and Wain-Hobson,S. (1990) DNA recombination during PCR. *Nucleic Acids Res.*, **18**, 1687–1691.
30. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
31. Fehlmann,T., Backes,C., Alles,J., Fischer,U., Hart,M., Kern,F., Langseth,H., Rounge,T., Umu,S.U., Kahraman,M. *et al.* (2017) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, **34**, 1621–1628
32. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database, C. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
33. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Wurstle,M.L., Hubenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.
34. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stahler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
35. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement.*, **12**, 565–576.
36. Graham,F.L., Smiley,J., Russell,W.C. and Nairn,R. (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J. Gen. Virol.*, **36**, 59–74.
37. Thomas,P. and Smart,T.G. (2005) HEK293 cell line: a vehicle for the expression of recombinant proteins. *J. Pharmacol. Toxicol. Methods*, **51**, 187–200.
38. Biasiolo,M., Sales,G., Lionetti,M., Agnelli,L., Todoerti,K., Bisognin,A., Coppe,A., Romualdi,C., Neri,A. and Bortoluzzi,S. (2011) Impact of host genes and strand selection on miRNA and miRNA* expression. *PLoS One*, **6**, e23854.
39. Hu,H.Y., Yan,Z., Xu,Y., Hu,H., Menzel,C., Zhou,Y.H., Chen,W. and Khaitovich,P. (2009) Sequence features associated with microRNA strand selection in humans and flies. *BMC Genomics*, **10**, 413.
40. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
41. Schamberger,A., Sarkadi,B. and Orban,T.I. (2012) Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner. *RNA Biol.*, **9**, 1177–1185.
42. Hube,F., Ulveling,D., Sureau,A., Forveille,S. and Francastel,C. (2017) Short intron-derived ncRNAs. *Nucleic Acids Res.*, **45**, 4768–4781.
43. Hansen,T.B., Veno,M.T., Jensen,T.I., Schaefer,A., Damgaard,C.K. and Kjems,J. (2016) Argonaute-associated short introns are a novel class of gene regulators. *Nat. Commun.*, **7**, 11538.
44. Magee,R., Telonis,A.G., Cherlin,T., Rigoutsos,I. and Londin,E. (2017) Assessment of isomiR discrimination using commercial qPCR methods. *Noncoding RNA*, **3**, 18.
45. Pillman,K.A., Goodall,G.J., Bracken,C.P. and Gantier,M.P. (2019) miRNA length variation during macrophage stimulation confounds the interpretation of results: implications for miRNA quantification by RT-qPCR. *RNA*, **25**, 232–238.
46. Nejad,C., Pepin,G., Behlke,M.A. and Gantier,M.P. (2018) Modified polyadenylation-based RT-qPCR increases selectivity of amplification of 3′-MicroRNA isoforms. *Front. Genet.*, **9**, 11.
47. Hou,S., Fang,M., Zhu,Q., Liu,Y., Liu,L. and Li,X. (2017) MicroRNA-939 governs vascular integrity and angiogenesis through targeting gamma-catenin in endothelial cells. *Biochem. Biophys. Res. Commun.*, **484**, 27–33.
48. Aghdaei,F.H., Soltani,B.M., Dokanehiifard,S., Mowla,S.J. and Soleimani,M. (2017) Overexpression of hsa-miR-939 follows by NGFR down-regulation and apoptosis reduction. *J. Biosci.*, **42**, 23–30.
49. Zhang,J.X., Xu,Y., Gao,Y., Chen,C., Zheng,Z.S., Yun,M., Weng,H.W., Xie,D. and Ye,S. (2017) Decreased expression of miR-939 contributes to chemoresistance and metastasis of gastric cancer via dysregulation of SLC34A2 and Raf/MEK/ERK pathway. *Mol. Cancer*, **16**, 18.
50. Chen,C., Wu,M., Zhang,W., Lu,W., Zhang,M., Zhang,Z., Zhang,X. and Yuan,Z. (2016) MicroRNA-939 restricts Hepatitis B virus by targeting Jmjd3-mediated and C/EBPalpha-coordinated chromatin remodeling. *Sci. Rep.*, **6**, 35974.
51. Guo,Z., Shao,L., Zheng,L., Du,Q., Li,P., John,B. and Geller,D.A. (2012) miRNA-939 regulates human inducible nitric oxide synthase posttranscriptional gene expression in human hepatocytes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 5826–5831.
52. Yang,F., Nam,S., Brown,C.E., Zhao,R., Starr,R., Ma,Y., Xie,J., Horne,D.A., Malkas,L.H., Jove,R. *et al.* (2014) A novel berbamine derivative inhibits cell viability and induces apoptosis in cancer stem-like cells of human glioblastoma, via up-regulation of miRNA-4284 and JNK/AP-1 signaling. *PLoS One*, **9**, e94443.
53. Li,Y., Shen,Z., Jiang,H., Lai,Z., Wang,Z., Jiang,K., Ye,Y. and Wang,S. (2018) MicroRNA4284 promotes gastric cancer tumorigenicity by targeting ten-eleven translocation 1. *Mol. Med. Rep.*, **17**, 6569–6575.
54. Kim,Y.K., Kim,B. and Kim,V.N. (2016) Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E1881–E1889.

3.8 Distribution of miRNA expression across human tissues

Published online 25 February 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 8* ***3865–3877***
doi: 10.1093/nar/gkw116

Distribution of miRNA expression across human tissues

Nicole Ludwig[1], Petra Leidinger[1], Kurt Becker[2], Christina Backes[3], Tobias Fehlmann[3], Christian Pallasch[4,5], Steffi Rheinheimer[1], Benjamin Meder[6,7,8], Cord Stähler[9], Eckart Meese[1] and Andreas Keller[3,*]

[1]Institute of Human Genetics, Saarland University, Medical School, Homburg, Germany, [2]Institute of Anatomy and Cell Biology, Saarland University, Medical School, Homburg, Germany, [3]Chair for Clinical Bioinformatics, Saarland University, Saarbruecken, Germany, [4]Department I of Internal Medicine and Center of Integrated Oncology, University Hospital of Cologne, Cologne, Germany, [5]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Cologne, Germany, [6]Department of Internal Medicine III, University Hospital Heidelberg , 69120 Heidelberg, Germany, [7]German Center for Cardiovascular Research (DZHK), 69120 Heidelberg, Germany, [8]Klaus Tschira Institute for Integrative Computational Cardiology, D-69118 Heidelberg, Germany and [9]Siemens Healthcare, Hartmannstrasse 16, 91052 Erlangen, Germany

Received May 06, 2015; Revised February 17, 2016; Accepted February 17, 2016

ABSTRACT

We present a human miRNA tissue atlas by determining the abundance of 1997 miRNAs in 61 tissue biopsies of different organs from two individuals collected post-mortem. One thousand three hundred sixty-four miRNAs were discovered in at least one tissue, 143 were present in each tissue. To define the distribution of miRNAs, we utilized a tissue specificity index (TSI). The majority of miRNAs (82.9%) fell in a middle TSI range i.e. were neither specific for single tissues (TSI > 0.85) nor housekeeping miRNAs (TSI < 0.5). Nonetheless, we observed many different miRNAs and miRNA families that were predominantly expressed in certain tissues. Clustering of miRNA abundances revealed that tissues like several areas of the brain clustered together. Considering -3p and -5p mature forms we observed miR-150 with different tissue specificity. Analysis of additional lung and prostate biopsies indicated that inter-organism variability was significantly lower than inter-organ variability. Tissue-specific differences between the miRNA patterns appeared not to be significantly altered by storage as shown for heart and lung tissue. MiRNAs TSI values of human tissues were significantly ($P = 10^{-8}$) correlated with those of rats; miRNAs that were highly abundant in certain human tissues were likewise abundant in according rat tissues. We implemented a web-based repository enabling scientists to access and browse the data (https://ccb-web.cs.uni-saarland.de/tissueatlas).

INTRODUCTION

Knowing the expression and distribution of different molecule classes in tissues is essential for the understanding of both physiological and pathological mechanisms. The gene expression atlas (1), hosted at the European Bioinformatics Institute, collects gene expression patterns under different biological conditions in various organisms. Likewise, the Human Protein Atlas presents information on proteomes in various tissues (2). For the class of small non-coding nucleic acids, the so-called microRNAs or miRNAs, there is a lack of up-to-date databases showing their tissue-specific distribution. The first and as of now most comprehensive analysis of miRNA abundance in different tissues has been reported by Landgraf et al. in 2007 (3). This sequencing-based study reported 340 miRNAs in 26 organs. We recently investigated the miRNA repertoire of different blood cell types (4), already indicating a complex miRNA repertoire strongly dependent on the considered cell types. To improve the understanding of the miRNA abundance in human tissues, we now profiled 1997 different mature miRNAs for 61 tissues. In contrast to the previous catalogue of miRNAs in human tissues, we measured all miRNA profiles from only two different individuals to minimize inter-individual variability. We selected an array-based analysis to have a robust platform for determining the miRNA abundance. The applied Agilent microarray technology has been proven sensitive and, more important, reproducible in a recent comprehensive platform comparison (5). Using this technology, we achieved technical Pearson correlation co-

[*]To whom correspondence should be addressed. Tel: +49 681 302 68611; Fax: +49 681 302 58094; Email: andreas.keller@ccb.uni-saarland.de

efficients of between 0.97 and 1 for technical replicates in previous studies.

Here, we first characterize technical stability of our approach before we describe variations in the abundance of the miRNAs across tissues. To provide easy access to the tissue atlas, we implemented a web-based repository that also links results to important miRNA resources. This web service is freely available online at https://ccb-web.cs.uni-saarland.de/tissueatlas.

MATERIALS AND METHODS

Tissues and RNA extraction

Tissues analysed in this study originated from two male bodies. Both cadavers were obtained as anatomical gift to be dissected in a study of medicine under German law. The first body was from a 65-year-old male patient, who suffered from multiple myeloma, a cancer that forms in a type of white blood cells (plasma cells). The body was stored at 4°C upon arrival at the anatomical institute and tissue samples were collected the following day, i.e. 2 days post-mortem. In total, we analysed 24 different tissues, i.e. adipocytes, arachnoid mater, artery, colon, small intestine (ileum), dura mater, brain, urinary bladder, skin, myocardium, bone (rib), liver, lung, stomach, spleen, muscle, gall bladder, muscle fascia, epididymis, intercostal nerve, kidney, thyroid, testis and tunica albuginea of testis.

The second body was from a 59-year-old male individual, who died a natural death. The body was frozen at −20°C after arrival at the anatomical institute and dissected after 3 weeks of storage. Autopsy showed no signs of cancer. As we aimed at increasing the resolution our tissue atlas, we collected 37 samples including several sub-areas for different organs, i.e. nine brain areas (grey matter, white matter, frontal, temporal, occipital, nucleus caudatus, thalamus, pituitary gland and cerebellum), dura mater, spinal cord, nerve, artery, vein, myocard, muscle, lymph node, thyroid, esophagus, stomach, pancreas, duodenum, jejunum, colon, liver, three kidney areas (kidney unspecified, medulla and cortex), spleen, adrenal gland, prostate, testis, skin, adipocyte, lung, pleura and bone marrow.

To assess the influence of RNA degradation originating from different storage times of the tissue on the miRNA profile, we used normal lung and normal heart tissue that was stored in physiological salt solution at 4°C for 1, 2, 3, 7 and 14 days, before RNA isolation. To understand short-term effects on the miRNA pattern in a comprehensive manner, we analysed lung tissue from another individual. The following 16 time points were profiled: 0, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 9, 12, 24, 36, 48, 72 and 96 h.

To estimate inter-individual variations, we exemplarily performed in-depth analysis for lung tissues. For 16 normal tissue biopsies from different individuals, the miRNA expression intensity was determined as for the two bodies and the samples from the degradation analysis.

RNA isolation and integrity

RNA was isolated using the miRNeasy Mini Kit (Qiagen) and the Qiagen tissue lyser using 7 mm stainless steel beads. Tissue samples were disrupted for 5 min 30 Hz (1800 oscillations/min) in Qiazol lysis reagent. Further purification was done according to manufacturer's instructions. Concentration and purity was measured using NanoDrop 2000 (Thermo Scientific). RNA integrity was measured using Bioanalyzer RNA Nano Chip (Agilent). As expected for autopsy samples, the RNA integrity values (RIN) ranged between 1.8 and 2.7.

miRNA profiling

Microarray analysis was performed using *SurePrint* 8 × 60K Human V19 *miRNA* microarrays (Agilent) that contain 2007 miRNAs of miRBase V19 (http://www.mirbase.org/), according to the manufacturer's instructions for the first corpse. For the second corpse, the most recent miRBAse v21 has been used and the analysis has been carried out on 1997 human miRNAs present in both versions. In brief, a total of 100 ng RNAs were processed using the miRNA Complete Labeling and Hyb Kit to generate fluorescently labelled miRNA. Microarrays were scanned with the Agilent Microarray Scanner at 3 μm in double path mode. Microarray scan data were further processed using Agilent Feature Extraction software. The raw expression intensity values are available for download at https://ccb-web.cs.uni-saarland.de/tissueatlas. Since the normalization may have an impact on the results, we performed all analyses on the raw data, normalized data by quantile normalization and by variance stabilizing normalization (6). For training the Variance Stabilized Normalization (VSN) model all samples and all miRNAs were used. The detailed results for the variance stabilizing normalization are provided in the supplementary material. To account for negative values (i.e. miRNAs that are not expressed, that may get a negative value due to background subtraction) a pseudo-count has been added. All calculations have been carried out in R version 3.0.2.

Tissue specificity index

To evaluate the variability of expression patterns, we calculated a tissue specificity index (TSI) for each miRNA analogously to the TSI 'tau' for mRNAs originally developed by Yanai et al. (7). This specificity index is a quantitative, graded scalar measure for the specificity of expression of a miRNA with respect to different organs. The values range from 0 to 1, with scores close to 0 represent miRNAs expressed in many or all tissues (i.e. housekeepers) and scores close to 1 miRNAs expressed in only one specific tissue (i.e. tissue-specific miRNAs). Specifically, the TSI for a miRNA j is calculated as

$$\mathrm{tsi}_j = \frac{\sum_{i=1}^{N}(1 - x_{j,i})}{N - 1},$$

where N corresponds to the total number of tissues measured and $x_{j,\mathrm{i}}$ is the expression intensity of tissue i normalized by the maximal expression of any tissue for miRNA j.

Hierarchical clustering of tissues

To estimate the proximity of profiles from different tissues, hierarchical clustering analysis has been applied. To account for the high dynamic range of miRNAs, clustering has been performed on log expression intensities and miRNAs that are close to the background were removed. To extend the cluster analysis, the 100 most variable miRNAs have been selected. In each case, complete linkage hierarchical clustering using the Euclidian distance has been performed.

Expression of miRNA families

For estimating the tissue specificity of miRNA families, we extracted all miRNA families from the most recent miRBase version 21. For each miRNA precursor all mature forms have been considered as family members, duplicated mature miRNAs (e.g. coming from different precursors in the same family) have been counted once in order to minimize a potential bias introduced by multiple precursors. For discovering co-expressed miRNAs, Spearman correlation of intensity values between all pairs of miRNAs has been calculated. Network visualization has been performed in Cytoscape.

Conservation of tissue specificity

To compare conserved tissue specificity in humans and rats, we downloaded data from the Gene Expression Omnibus (GEO) series GSE52754, containing expression profiles for 55 different rat tissues that have been measured using Agilent microarrays (8). To match miRNAs we extracted all rat miRNA identifiers from the respective manuscript and matched them via a 100% sequence match. For matching miRNAs and matching tissues, we calculated and correlated the tissue specificity indices. To minimize artefacts introduced by normalization, we carried out all analyses on raw data. Since this analysis only addresses the question whether a miRNA is rather specific or a housekeeping miRNA, we also correlated the human and rat expression profiles using Spearman correlation.

Additional data from literature

In addition to the 44 tissue samples from the degradation and reproducibility analysis, the 16 individual lung cancer tissues and the 61 tissues from two bodies newly measured for this study, we searched the literature for other studies where normal tissues have been profiled. In the GEO (9), we found 1178 series related to miRNAs. Of these, 722 were from *Homo sapiens*. Excluding series with low sample count (below 20 samples), 302 series remained. After excluding studies from body fluids such as serum, plasma, blood or urine, we examined the remaining hits for availability of unaffected tissue measurements. The respective data tables were downloaded from GEO and all IDs were matched from the respective platform identifiers to miRBase Version 21 IDs. For the respective studies, raw and normalized data (VSN and quantile normalized) were added to our tissue atlas web repository. These include 43 samples from 9 tissues and 463 miRNAs from GSE11879, 40 samples measured for 709 miRNAs from normal gastric tissue from GSE23739, 48 benign prostate tissues measured for 480 miRNAs from series GSE54516 and 32 benign prostate tissues measured for 825 miRNAs from series GSE76260. The data have been used partially in the present manuscript, all data are included in the web-based tissue atlas resource.

RESULTS

In this work, we present the draft of a human tissue miRNome atlas. In the first part of the manuscript, we describe pre-analytics, investigating the general reproducibility of the miRNA profile measurements and also the effect of storage of tissues on miRNA profiles. In the pre-analytics consideration, we measured 44 tissue miRNomes. It is essential to understand respective variability to understand the biological variability of different tissue miRNomes.

In the second part, we describe the screening of all mature miRNAs from miRBase version 21 across different organs of two male bodies. We investigated miRNA expression in 24 different tissues from the first body and in 37 different tissues from the second body. To determine the miRNAs abundance in the different tissues, we utilized a TSI score, known from transcriptomics (7). Furthermore, we investigated the proximity of organs based on miRNA abundances by hierarchical clustering and co-expression analysis. To estimate inter-individual variations, we measured 16 additional miRNomes from control lung tissues and extracted further data sets from the GEO.

To provide researchers access to the first version of the miRNA tissue atlas, we implemented a web-based repository that is freely available at www.ccb.uni-saarland.de/tissueatlas.

Reproducibility of miRNA patterns

An important factor for estimating the biological variability is to understand the technical variability of the underlying profiling platform. Previously, we compared technical reproducibility of the two common platforms, microarrays (Agilent) and NGS (Illumina HiSeq) (10). In these experiments, we discovered an increased variability of miRNAs dependent of the sequencing library preparation. Similarly, we observed a strong bias based on the nucleotide composition of miRNAs (11). Of 10 replicated Agilent microarray measurements of the same individual, we calculated 10 * 9/2 pair-wise correlations of technical replicates. Minimal correlation was 0.998 and mean/median correlation 0.999, highlighting the high degree of technical reproducibility of the array platform. To translate these results on our tissue atlas and determine technical reproducibility of the array analysis, technical duplicates from nine randomly selected tissue samples from the second body were measured. The duplicates were processed at different days and have been measured on different arrays, each. Hierarchical cluster analysis shows that the technical replicates always clustered together showing that the applied technology was suited to provide reproducible results (Figure 1 shows the heat map for quantile-normalized data, Supplementary Figure S1 for VSN-normalized data). Altogether, we found high correlations between these technical replicates with the overall lowest correlations at 0.986 and 0.994 observed for liver tissue and pleura, respectively. Highest correlation of 0.999 was reached for the brain samples.

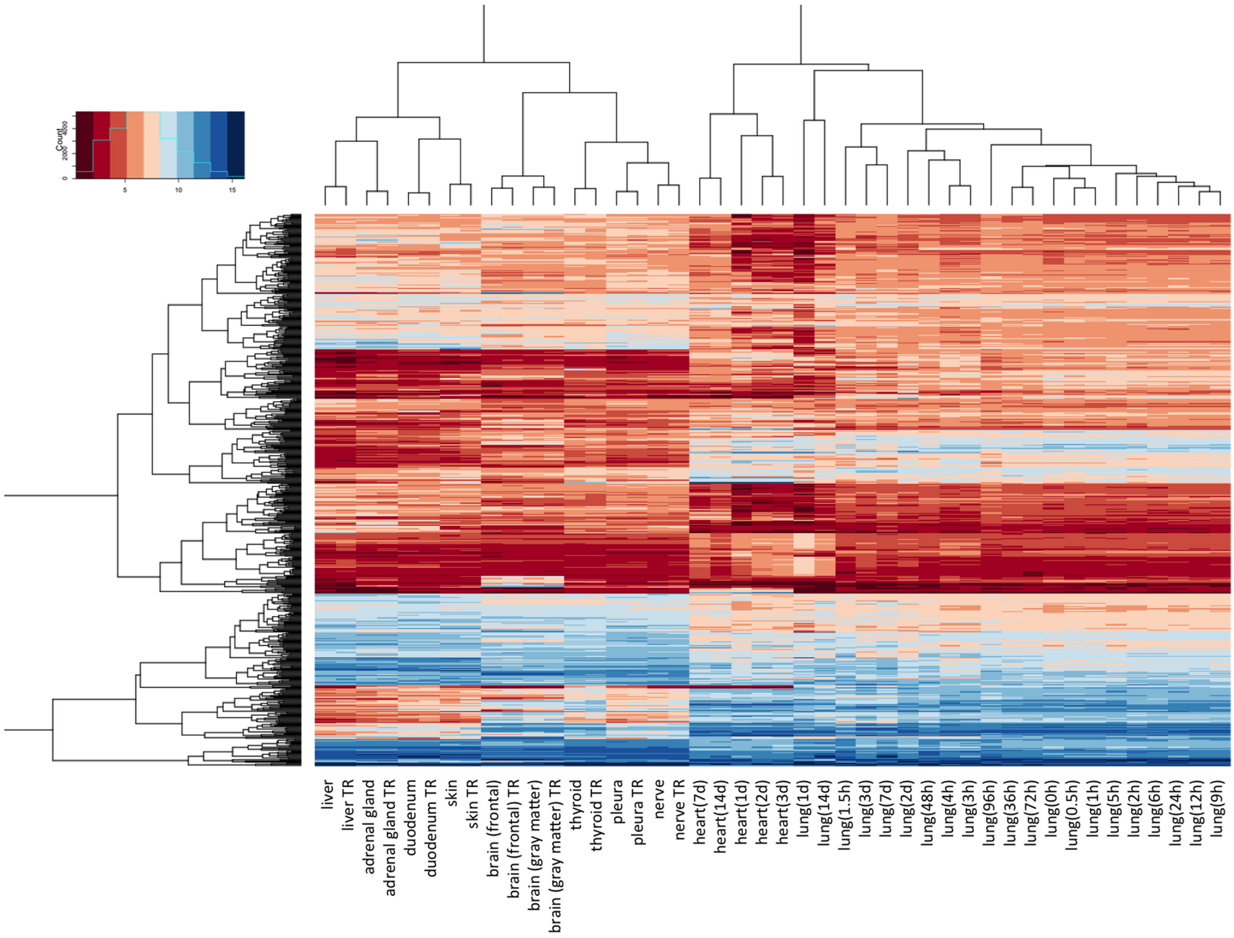


Figure 1. Hierarchical clustering of the 44 samples included in the stability and reproducibility study. Quantile normalized and $\log_2$ transformed expression intensity values were used for clustering. The intensity values and distribution are presented in the upper left corner. In the present heat map, heart and lung tissues cluster together on the right-hand side. Technical replicates (marked by 'TR' in the labels below the heat map) of other organs cluster together in each case in the left-hand side. For VSN-normalized data the same representation is provided in Supplementary Figure S1.

Stability of miRNA patterns in tissues

Measuring tissues of corpses the storage time prior to RNA extraction and a potential degradation of RNA may have an influence on the profiles. We exemplarily investigated the process for heart and lung tissue. Biopsies were taken from two individuals and have been stored for 1, 2, 3, 7 and 14 days at 4°C. Hierarchical cluster analysis shows that all lung and all heart samples each cluster together (Figure 1; Supplementary Figure S1). The duration of the storage was, however, not reflected in the clustering pattern indicating that a storage time between 1 and 14 days at 4°C has a limited influence on the overall miRNA tissue pattern.

We also performed the analysis with more dense time intervals within the first 3 days to understand short-term effects. For a lung tissue from a third individual 16 time points between 0 and 96 h were profiled. These biopsies clustered well together with the lung tissues from the second individual with storage time over 14 days. Again, no time curse could be recognized in the clustering pattern.

Remarkably, the results presented above describe the overall miRNA patterns. For single miRNAs still differences dependent on the storage could be observed. Thus, we calculated the TSI for all lung tissues and for all tissues in the pre-analytical study. With respect to lung tissues, large TSI values mean in this case not tissue specific but rather specific in one of the replicated measurements. We thus expect that TSI values of miRNAs from the lung tissue are low. Especially for five miRNAs we, however, calculated TSI values that are increased in lung tissue by at least 20%: hsa-miR-8069, hsa-miR-6821–5p, hsa-miR-4800–5p, hsa-miR-6775–5p, hsa-miR-5001–5p. For all miRNAs, TSI values from the pre-analytical step are summarized in Supplementary Table S1.

Frequency of miRNAs per tissue and tissue specificity of miRNAs

For each miRNA in each tissue, we determined its presence and frequency using the so-called present calls as determined by Agilent Feature Extraction software. Out of the 1997 different mature miRNAs, 633 (31.7%) were not detected in any of the tested tissues by the applied microarray technology. Out of the remaining 1364 miRNAs, 143 (10.5%) were found in all tissues. To present more comprehensive information on the tissue distribution of miRNAs, we utilized the miRNA TSI analogously to the mRNA TSI 'tau' that has successfully been employed by Yanai et al. (7). This index has a range of 0–1 with the score of 0 corresponding to ubiquitously expressed miRNAs (i.e. 'housekeepers') and a score of 1 for miRNAs that are expressed in a single tissue (i.e. 'tissue-specific' miRNAs). We calculated TSI for the 1364 miRNAs that have been detected in at least one tissue sample. For each miRNA, we compared TSI for the two bodies, for raw, quantile- and VSN-normalized data (Supplementary Table S2). Using the quantile-normalized data for the first body, 83.7% of all miRNAs showed an average abundance throughout the tissues with intermediate TSI values ranging from 0.15 to 0.85 (Figure 2A, Supplementary Figure S2A for VSN-normalized data). Only one miRNA (miR-3960) was ubiquitously expressed with a TSI < 0.15 and 222 miRNAs showed a highly tissue-specific expression with TSI > 0.85. For the second body, 88.8% of all miRNAs showed intermediate TSI values; one miRNA (miR-6089) showed a TSI < 0.15 and 152 miRNAs a TSI > 0.85 (Figure 2B, Supplementary Figure S2B for VSN-normalized data). The correlation of the VSN-normalized TSI values with the quantile-normalized TSI values was 0.88 ($P < 10^{-10}$).

The overall most tissue-specific miR-1–3p is presented in Figure 3. For all 61 samples raw-, quantile- and VSN-normalized expression intensities are presented as bar plot. Respective bar plots for all miRNAs can be generated using the online repository.

Clustering of tissue patterns and analysis of miRNA families

Beyond the analysis of single miRNAs, we determined the overall similarity/dissimilarity of the miRNA pattern between the different tissues. We performed hierarchical clustering of miRNAs and tissues using normalized expression intensities. We found two major clusters, the first of which containing mainly nervous system tissues and muscle tissues from both bodies. In the second cluster, the organs of the two individuals frequently did not cluster together (Figure 4A). Since the large number of miRNAs used for this clustering likely caused substantial noise, we restricted the clustering analysis to the 100 miRNAs with the highest data variance (Figure 4B). Here, we found three main clusters with the first one containing kidney, liver, stomach and small intestine of both bodies. The second cluster exclusively contained all brain tissue samples of both bodies and nervous system related tissue, i.e. spinal cord and dura mater. The third cluster contained thyroid, nerve, muscle, myocardium and colon each of both bodies. Other organs were found in different clusters, e.g. the lung samples and the brain coverings dura mater and arachnoid mater. For VSN-normalized data we observed a similar pattern, however, we found a stronger tendency of clustering of individuals in the different sub-clusters (Supplementary Figure S3).

To gain further insights into expression of tissue-specific miRNAs, we performed clustering with the 25 miRNAs displaying a TSI $>$ 0.85 for both bodies in raw-, quantile- and VSN-normalized data (Figure 5). We found several groups of miRNAs with tissue-specific expression. In detail, we detected high expression of miR-133b, miR-133a-3p, miR-1–3p and miR-206 in both muscle samples and, with the exception of miR-206 also in both myocardial samples. Additionally, we found a cluster of four miRNAs specifically expressed in various brain tissues, i.e. miR-338–3p, miR-219a-5p, miR-124–3p and miR-9–5p. Another group of miRNAs, miR-507, miR-514a-3p and miR-509–5p was almost exclusively expressed in the testis samples. Besides these miRNA clusters, we also found single miRNAs that were expressed in a highly tissue-specific manner, i.e. miR-122–5p, miR-7–5p and miR-205–5p were each exclusively expressed in liver, pituitary gland and skin, respectively.

Tissue specificity of miRNA families

To further determine to what extend miRNA families show similar abundances in specific organs, we calculated the TSI not only for single miRNAs but also for mature miRNAs inside each miRNA family. Out of 187 miRNA families from the miRBase with at least two family members, we analysed 25 miRNA families with at least five mature forms (Figure 6A; Supplementary Table S3). We found several miRNA families with high TSI values including the above-mentioned mir-378 family with most of the family members showing a high abundance in muscle tissues and the myocardium. Similarly, the mir-506 family with 18 family members showed generally a high abundance in testis tissue while they were less expressed in other tissues. Other families, such as the mir-449 family with five members, did not show a common pattern in the different tissues: MiR-449c-3p was expressed specifically in spleen tissue, miR-449c-5p and -449b-5p in kidney and small intestine, miR-449a in lung, kidney and brain and miR-449b-3p in spleen. To extend this analysis we searched for miRNAs co-expression patterns in specific tissues. We used a high correlation cut-off and considered only miRNA-pairs with Pearson correlation exceeding 0.95. Altogether, we identified 73 miRNA pairs with tissue co-expression. In addition to pair-wise interactions, we also found sub-networks with at least four participants. The networks have been visualized using Cyto-Scape (Figure 6B). While we frequently observed co-expression among mature members of specific families (e.g. the mir-548 family), we also found correlations of miRNAs from different miRNA families. For example, miR-4312 was co-expressed with miRNAs from the let-7 family. Performing the same analysis with raw data, we detected an increased number of co-expressions, but generally confirmed the observation that has been based on the normalized data.

Tissue specificity of -3p and -5p mature forms

We asked whether -3p and -5p mature forms of miRNAs have different tissue specificity. To limit the bias of miR-

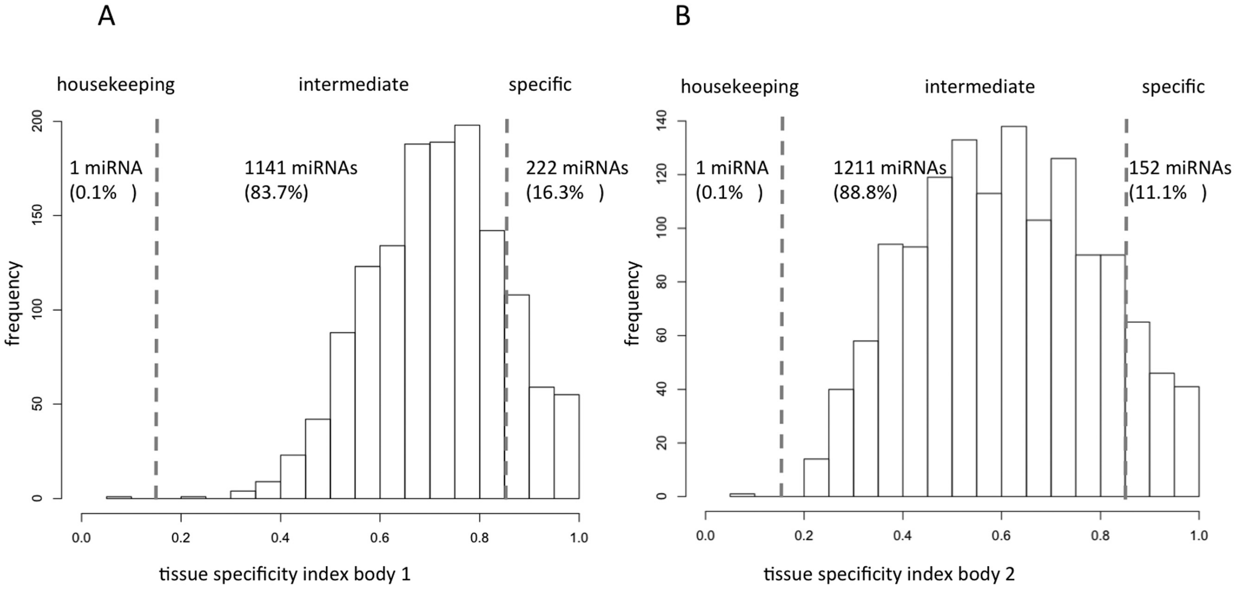


Figure 2. Histogram plot for the frequency of TSI of miRNAs in different tissues. Panel **A** represents TSI of the first, panel **B** of the second body. The vertical dotted lines correspond to the threshold originally proposed for defining housekeeping and specifically expressed miRNAs of <0.15 and >0.85. The same representation for VSN-normalized data is presented in Supplementary Figure S2.

NAs that are annotated with only one mature form, we only included those miRNAs that have two mature forms annotated and carried out the analyses in a paired manner (41% of the 1364 mature miRNAs were included). First, we investigated whether -3p or -5p mature forms are overall higher expressed. For both quantile- and VSN-normalized data, we calculated significantly higher expression of the -5p mature forms. The effects in VSN exceeded the quantile-normalized effects. Mature -5p forms were on average 21% higher expressed as compared to -3p forms (paired t-test *P*-value of 3.6×10^{-10}). To estimate whether the two mature forms are more or less specific for tissues, we calculated and compared the TSI values for the -3p and -5p forms. For both, TSI values based on VSN- and quantile-normalized data, we did not found significant differences between -3p and -5p forms ($P > 0.5$ in both cases). Having a detailed look at single miRNAs, we discovered that in all cases where -3p and -5p mature forms were tissue specific independent on the normalization technique the tissue patterns matched. The best matching profiles were found for hsa-miR-140, hsa-miR-378a, hsa-miR-509, hsa-miR-122, hsa-miR-124, hsa-miR-192 and hsa-miR-455. Only for one miRNA, miR-150, no significant correlation for -5p and -3p mature form was calculated (Supplementary Figure S4). The -3p form was specific for pancreas and the -5p form for stomach. All TSI values for -3p and -5p mature forms of quantile- and VSN-normalized data are available in Supplementary Table S4.

Inter-individual variations

In the previous analyses, we suggested that miRNAs are tissue specific. From two bodies it is impossible to extrapolate inter-individual variations within specific organs. In a first approach we searched for miRNAs that are overall higher or lower in all tissues of one of the two bodies, independent of the normalization technique. Two miRNAs, hsa-miR-548n and hsa-miR-548ap-5p, fulfilled these stringent criteria. Although these (and similarly differentially abundant miRNAs between both individuals) miRNAs had low TSI values and are not considered tissue specific the differences emphasize the importance of incorporating inter-individual variations.

We exemplarily analysed 16 lung tissue biopsies of 16 different individuals. Here, we expect miRNAs to be more homogenously expressed, leading to overall lower TSI values. For the quantile- and VSN-normalized data, we calculated significantly decreased TSI values in the individuals ($P < 10^{-16}$). The respective TSI values for biological replicates of lung tissue and the two bodies are presented in Supplementary Figure S5A (quantile normalized) and 5B (VSN normalized). These figures also indicate that few miRNAs have higher TSI in lung as compared to the overall TSI, i.e. variations between organs are smaller than variations between individuals. Inspecting the respective miRNAs, we found that they usually were specific for other organs than the lung and expressed to a very moderate limit in the lung. Here, already small variations lead to artificially high TSI values.

As the second example we downloaded expression values from 32 prostate tissues from the GEO (not affected tissues as part of a case-control cancer study, GSE76260). The TSI values were calculated for quantile- and VSN-normalized intensity values. Only the 625 miRNAs that were included in both studies were considered. In this analysis the variations between individuals were even lower as compared to

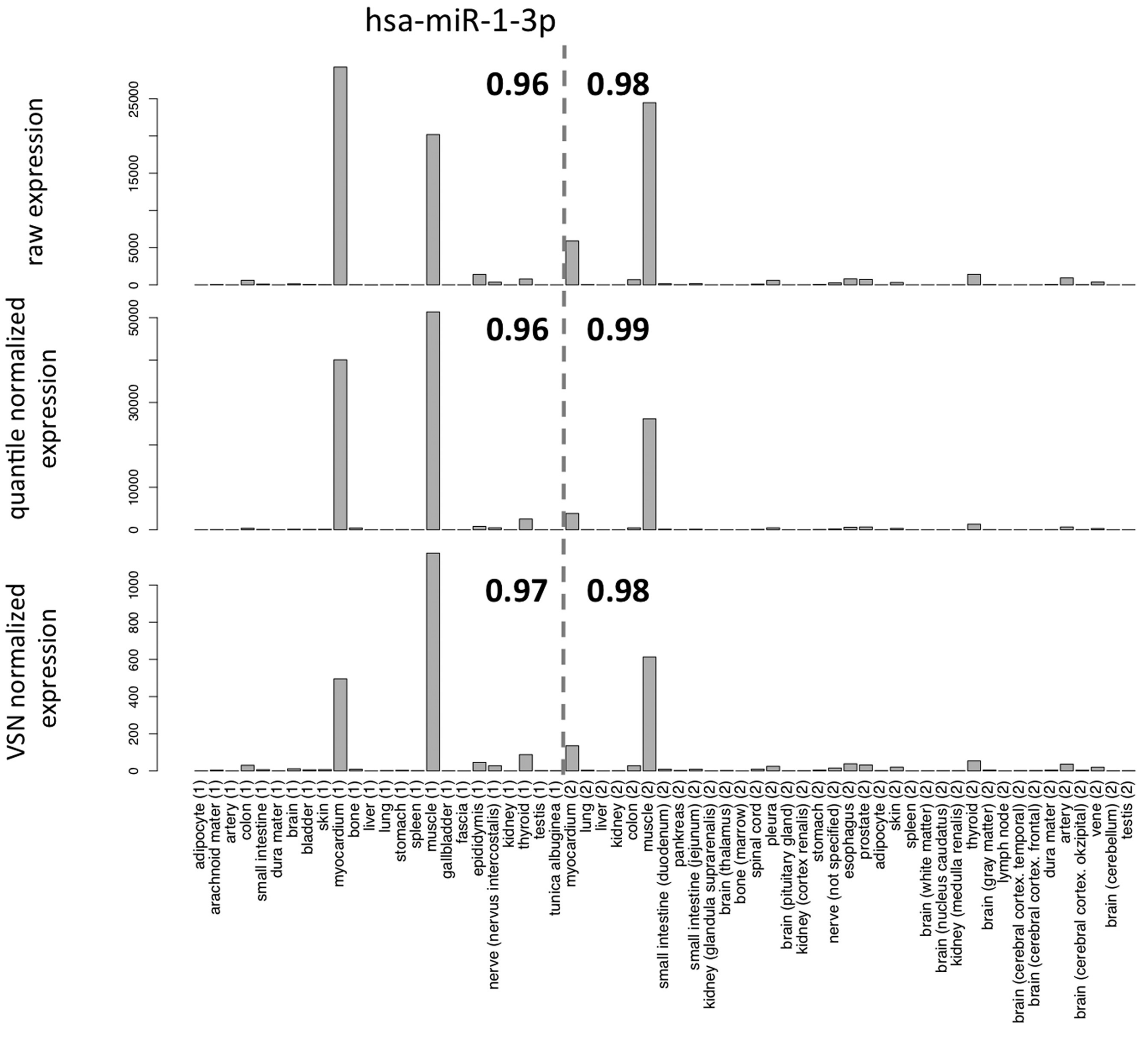


Figure 3. Bar plots for all 61 samples for miR-1–3p, the miRNA with highest overall TSI in the first and second body. The vertical dashed line separates the first from the second body. TSI values for both bodies are highlighted in the figure. The miRNA is high expressed in muscle and myocardium. Raw-, quantile- and VSN-normalized expression intensities for this miRNA match well across all different tissues.

the variations between organs. Again, TSI values were significantly lower for prostate tissue ($P < 10^{-16}$). The scatter plots are analogously to the lung tissues presented in Supplementary Figure S6. Also for the other tissues extracted from the GEO, which are also available on the tissue atlas web resource, lower TSI values were observed. In sum our results thus indicate that the inter-individual variations are smaller as compared to inter-organ variability.

Homology of tissue specificity in humans and rats

To addressed the question to what extend a tissue-specific abundance of the miRNA pattern is conserved between human and rodents, we matched the data of our study to data published in a recent study, which used the same miRNA platform (Agilent) (8). From all miRNAs expressed in our tissue collection, 230 matched in sequence identically between human and rat. Of the tissues included in the human and rat studies, 42 organs could be matched. For all these miRNAs and organs, we calculated the TSI values in human and rat, showing an overall correlation of 0.362 (P-value of 9×10^{-8}). To determine the significance of this finding, we additionally performed 1 million permutation tests, which showed an average correlation value of 0. While these results indicate an overall matching of miRNA abundances in humans and rats, the TSI does not acknowledge the origin of the miRNAs, i.e. a value of 1 for a rat miRNA may indicate

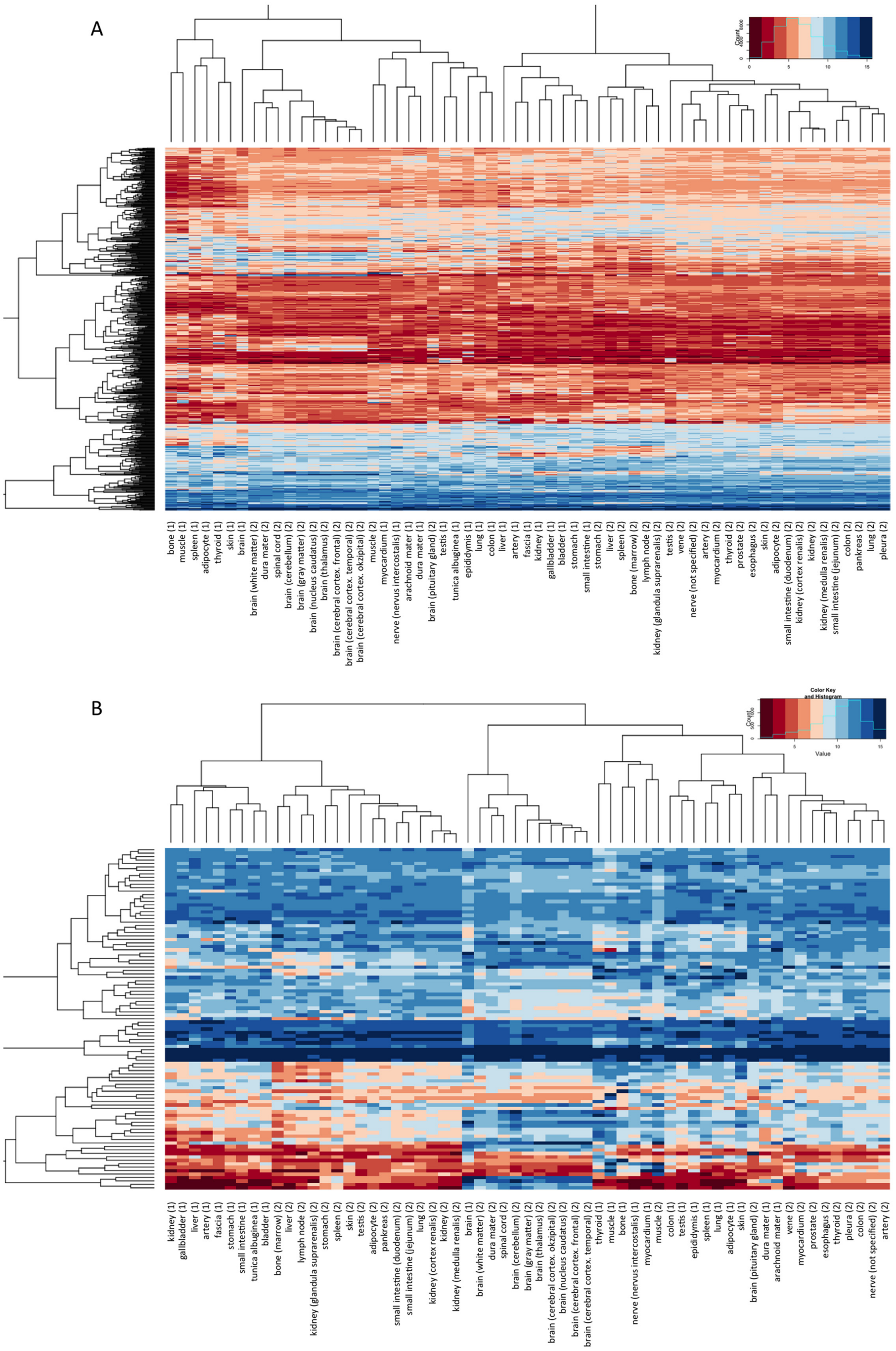


Figure 4. Hierarchical clustering of all tissues in both bodies. Log_2 transformed quantile normalized intensity values were used for clustering. The intensity value distribution is shown in the upper right corner of the figures. Panel **A** shows significantly expressed miRNAs, while panel **B** focuses on the 100 miRNAs with overall highest data variance. The respective representation for VSN-normalized data is presented in Supplementary Figure S3.

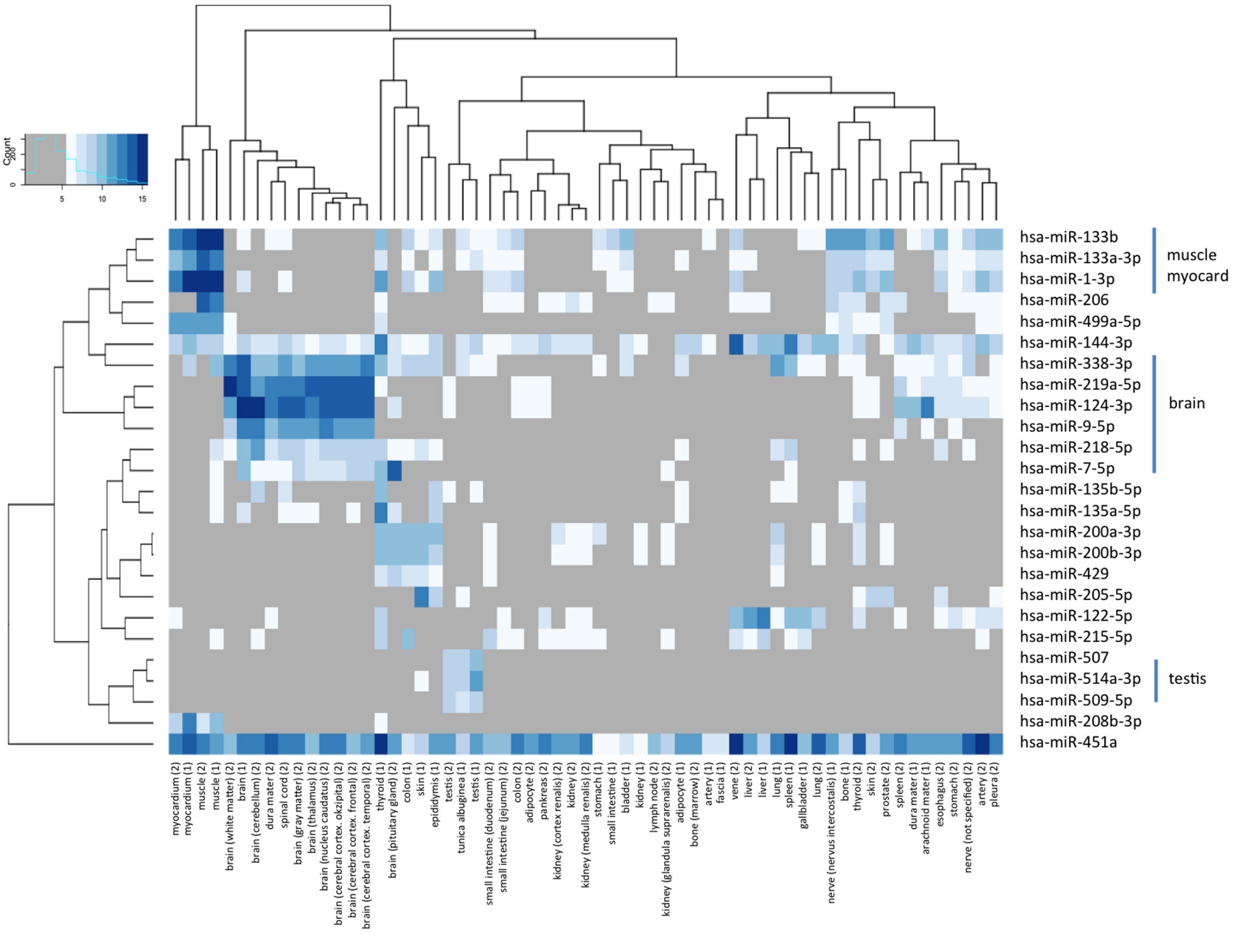


Figure 5. Heat map for the 25 miRNAs that have TSI values of >0.85 in both bodies. Log_2 transformed expression intensities of quantile normalized expression values are presented. To facilitate the interpretation of specific miRNAs in organs or organ groups low expressed miRNAs were greyed out (see also color distribution scheme in the upper left corner). The analysis highlights tissue-specific miRNAs that are exemplarily presented on the right-hand side of the plot, such as hsa-miR-1–3p that has already been described in Figure 3 as most specific miRNA overall.

specificity for spleen and for the same miRNA specificity for brain in humans. However, the overall correlation of the expression values of rat and human miRNAs was 0.361 ($P < 10^{-16}$), indicative of a significant matching of human and rat expression profiles. Similar to the results for humans in Figure 5, we clustered the miRNAs with high TSI values in human and rat. Altogether, we focused on very specific miRNAs: 54 miRNAs with TSI values exceeding 0.9 were considered. The resulting heat map where maximal rat and human miRNA expression was set to 100% to make both data sets comparable to each other is presented in Figure 7. In this analysis we did not observe a predominant clustering in humans and rats but a strong tendency of organs to cluster together. Examples of directly matching pairs include the spleen, myocardium, muscle, pancreas, kidney, liver, stomach, skin, brain or spinal cord. The miRNAs in this heat map matched the specific miRNAs in Figure 5 very well such as miR-133a-3p, and miR-133b for muscle and myocardium or miR-9–5p, miR-219a-5p, miR-7–5p and miR-124–3p for brain and spinal cord. Bar plots comparing each miRNA directly for specificity in tissues of rat and human are provided in the supplementary material.

DISCUSSION

As miRNAs emerge as important regulators of protein expression during tissue development and homeostasis, there is an increasing need for a standardized atlas of miRNA expression in multiple human tissues. Although there is ample evidence for differential miRNA expression in different human tissues, the majority of studies investigate differential expression in only one organ/tissue. Due to the different identification methods and normalization strategies, the results of these studies are not easily comparable limiting their value for comparison of miRNA expression in different tissues. The optimal human miRNA tissue atlas would be based on different fresh tissues each obtained from the same donor; different donors should be of different age and gender both of which are known to influence the miRNA

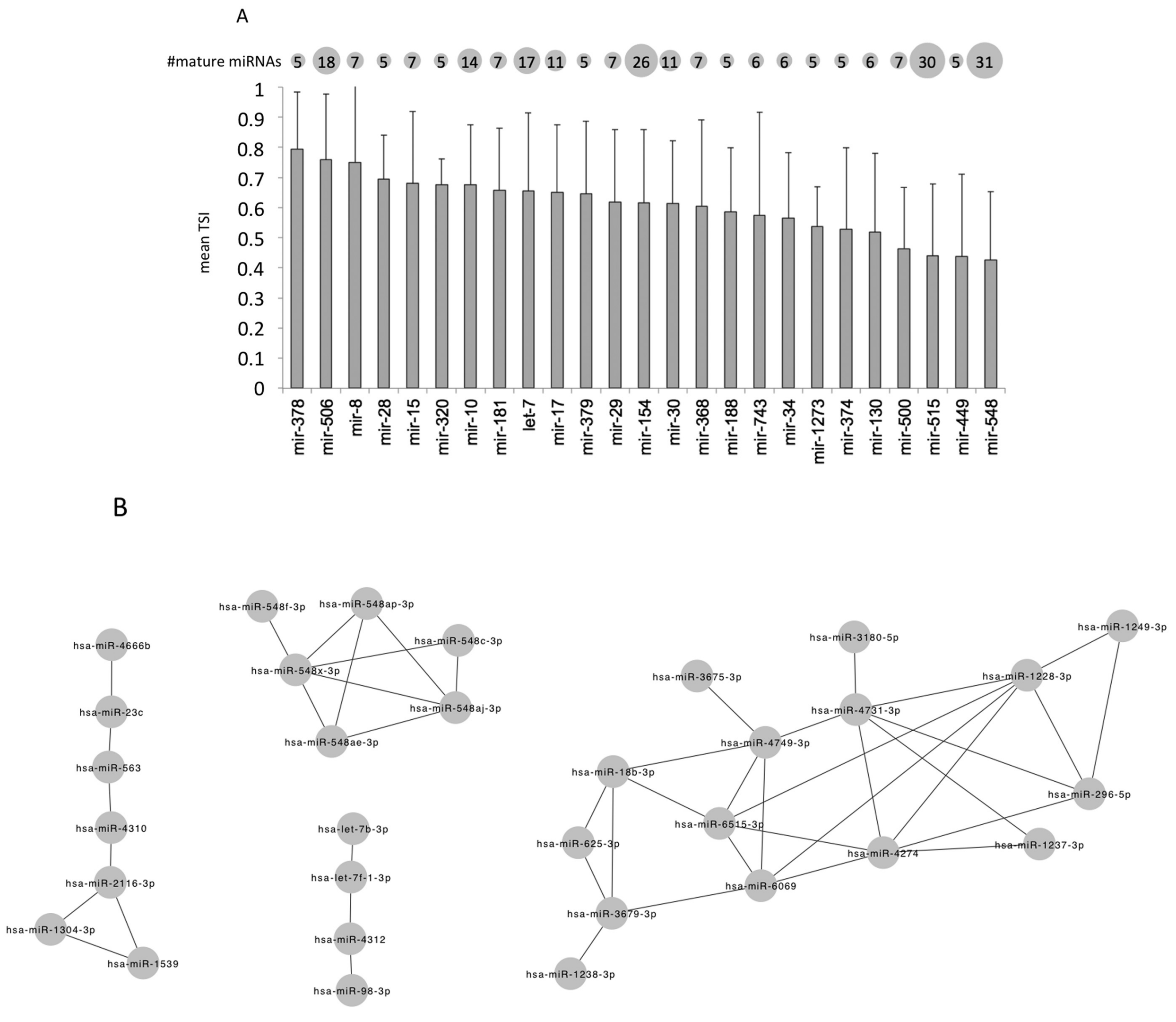


Figure 6. **A**: Average and standard deviation of TSI value in different miRNA families. For each miRNA family with at least five members the mean and standard deviation of all family members TSI is presented as bar plot. Families are sorted with decreasing average tissue specificity from left to right. Highest tissue specificity was observed for the miR-378 family, predominantly being specific for myocardium and muscle. The number of mature family members is shown above the columns with balloons, representing the family size. **B**: Co-expression network of miRNAs. Each miRNA pair connected by an edge has co-expression across all samples with Spearman correlation coefficient above 0.95.

pattern (12). As this ideal scenario is not possible in human studies, fresh biopsy material could be used for miRNA isolation with the advantage of yielding high-quality RNA. There are, however, several disadvantages: (i) biopsies will be mostly taken from patients with affected organs, (ii) high inter-individual differences can mask tissue-specific differences of miRNA abundances, (iii) a bias is likely introduced by multiple centres that are involved in tissue collections and (iv) samples of vital organs, e.g. thalamus, spinal cord or cerebellum, are not available. Alternatively, miRNAs can be isolated from tissues collected from the same individuals upon autopsy. The advantage of the latter approach is the availability of multiple tissues from the same individuals, even from vital organs, with the disadvantage of RNA degradation in the samples due to the storage duration of the body and the advanced age or the disease status of the body donors. In context of our tissue atlas, the main question is whether the differences in the abundance of miRNAs induced by post-mortem RNA degradation, which is different from *in-vitro* RNA degradation by UV light or heat, are higher than the differences between the tissues profiled. There is scant evidence for extended post-mortem stability of individual miRNAs (13,14). In case of whole miRNA tissue profiles, Ibberson et al. found that RNA degradation due to prolonged inadequate tissue storage has a random effect on miRNAs and compromises the reliability of miRNA

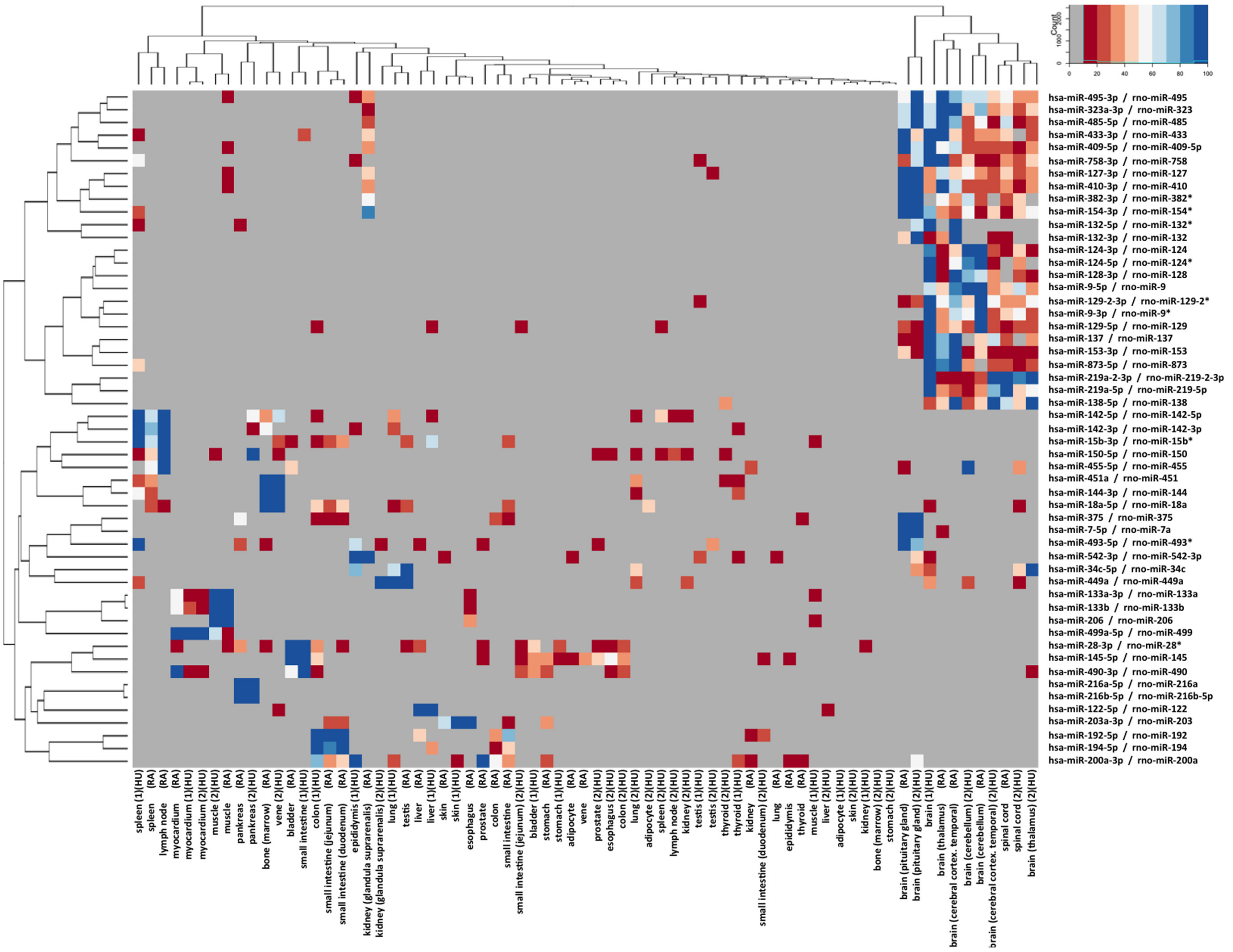


Figure 7. Conservation of tissue-specific expression of miRNAs in human and rat. Matching miRNAs (100% matching of mature miRNA sequence) from organ expression in rats and humans were calculated. For each miRNA in rats and humans the TSI was calculated and highly specific miRNAs were clustered. Since overall expression in humans and rats varied, the maximal intensity of each miRNA in the two organs was set to 100% and all other miRNAs were linearly scaled. All miRNAs with below 10% expression of maximal intensity are shown in grey to facilitate data interpretation (see also colour gradient presented in the upper right corner). On the right-hand side the human/rat miRNA identifiers are shown, below the heat map the matched tissues are presented (HU for human; RA for rat). For rat tissues the average intensity of replicated measurements is presented.

profiles, generating false positive deregulated miRNAs (15). But they also clearly state that 'even samples with the most degraded RNAs still preserve a tissue-specific miRNA signature'. This finding is in line with our observations in the present study. For lung and heart tissue we investigated short- and long-term degradation, highlighting an overall limited impact on the tissue specificity of miRNA profiles. Only very few miRNAs were affected at all. Given the data from two organs, we however cannot exclude the possibility that some tissue-specific miRNAs might be affected by degradation of the sample. We are also aware that the autopsy samples of the two male individuals provide only a snapshot of the full variability of miRNA expression. While we aim at adding more full body profiles we supported the data in the present study by tissue collections extracted from the literature (e.g. gastric and prostate tissues) and by own measurements (lung tissue).

We used a microarray platform for miRNA expression detection since this platform shows a high reproducibility as evidenced by the miRQC study (5). In our study, analysis of technical replicates of nine samples processed in different batches reached high correlation values above 0.986 for all samples. In previous studies, we observed a substantial bias introduced in Next Generation Sequencing (NGS) data by sample preparation of blood samples (10). However, NGS analysis would enable to detect presently unknown miRNAs as well miRNAs iso-forms that have demonstrated to target biological pathways in a cooperative manner (16). A key challenge with microarray data is normalization. Many techniques that are frequently applied such as variance stabilizing normalization or quantile normalization can have a substantial influence on the results. Quantile normalization e.g. assumes an overall similar distribution of all miRNAs. We thus performed the relevant analyses on raw data, quantile- and VSN normalization. Irrespective of the normalization technique we found higher TSI values for miRNAs as, e.g. known from mRNAs (7). This result suggests that miRNA expression is more tissue specific as compared to mRNA expression.

The, as of now, most comprehensive study on tissue-specific miRNAs in humans was published by Landgraf et al. in 2007 (3). They sequenced 256 small RNA libraries from 26 different organ systems and cell types of humans and rodents, with ∼1000 clone each. The human samples included normal samples from 16 tissues most of them brain and reproductive tissues. They identified 340 mature human miRNAs including 33 novel miRNAs not listed in the miRBase version 9.1, which was the current version at the time of the study (17). For canonical miRNAs they found a high concordance of tissue-specific expression in humans and rodents. When we compared our data to a data set on 55 different rat tissues available at GEO database (8), we could confirm conserved tissue-specific expression of several miRNAs, including miR-133b, miR-124 and miR-9. Amongst others, Landgraf et al. detected tissue-specific expression of miR-122 in liver, of miR-9, miR-124 and miR 128a/b in brain, of miR-7, miR-375, miR-141 and miR-200a in pituitary gland and of miR-142, miR-144, miR-150, miR-155 and miR-223 in hematopoietic cells. Overall, our results correlated well with this data, confirming specific expression of miR-122, miR-9, miR-124 and miR-7 in the respective organs. Consistent with Landgraf's results, we found miR-122–5p as highest expressed miRNA in the liver of both bodies. Our study, however, also identified low expression of miR-122–5p in spleen, gall bladder and veins. MiR-124 (miR-124–3p) was identified as the third most specific miRNA in the nervous system by Landgraf et al. We observed expression of this miRNA in different areas of the brain but not in other tissues. For miR-144, we found highest expression in vein and spleen, consistent with the assumption of residual hematopoietic cells in these samples; additionally, we found high expression of this miRNA in thyroid. Of note, miR-144 has been found highly expressed in normal thyroid and downregulated in papillary thyroid carcinoma (18). We also found high expression of miR-1–3p, miR-133a-3p, miR-133b and miR-206 in myocard and muscle. These miRNAs are known as myomiRs that regulate key genes in muscle development (19,20). Additionally, we detected a highly specific expression of miR-205–5p, miR-514a-3p and miR-192–5p in skin, testis and colon samples of one of the bodies, respectively. MiR-205–5p that is highly expressed in melanocytes and downregulated in melanoma is inverse correlated with melanoma progression (21). MiR-514a-3p belongs to the miR-506 family; the mouse orthologue of miR-506, mmu-201, has been shown to be specifically expressed in reproductive tissues (3). A significant decrease in expression of miR-192–5p in colorectal cancer compared to normal mucosa has been reported (22).

The knowledge of the expression pattern of miRNAs in different tissues is essential for understanding normal development and disease development of the respective tissue. In addition, knowing the tissues that express specific miRNAs helps to develop a miRNA found in whole blood or serum into a biomarker for a specific disease. Elevated serum levels of liver-specific miR-122 have been detected in patients with drug induced liver injury, steatosis, hepatitis-B and -C infections and in patients with hepatocellular carcinoma (23–26). Elevated levels of circulating myomiRs, i.e. miR-1, miR-206 and miR-133a/b, have been proposed as biomarker for heart failure and different forms of muscle dystrophy, but are also elevated after half-marathon run (27–29).

In summary, we provide an atlas of miRNA expression in multiple human tissues. This atlas can be used as starting point for elucidation of the role of miRNAs in tissue development and tissue-specific diseases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We acknowledge the support of Siemens Healthcare.

FUNDING

Saarland University and Siemens Healthcare. Funding for open access charge: Saarland University and Siemens Healthcare; funded in part by FP7 project BestAgeing.
Conflict of interest statement. None declared.

REFERENCES

1. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
2. Ponten,F., Jirstrom,K. and Uhlen,M. (2008) The human protein atlas–a tool for pathology. *J. Pathol.*, **216**, 387–393.
3. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
4. Leidinger,P., Backes,C., Meder,B., Meese,E. and Keller,A. (2014) The human miRNA repertoire of different blood compounds. *BMC Genomics*, **15**, 474.
5. Mestdagh,P., Hartmann,N., Baeriswyl,L., Andreasen,D., Bernard,N., Chen,C., Cheo,D., D'Andrade,P., DeMayo,M., Dennis,L. *et al.* (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.
6. Huber,W., von Heydebreck,A., Sultmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl. 1), S96–S104.
7. Yanai,I., Benjamin,H., Shmoish,M., Chalifa-Caspi,V., Shklar,M., Ophir,R., Bar-Even,A., Horn-Saban,S., Safran,M., Domany,E. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.
8. Minami,K., Uehara,T., Morikawa,Y., Omura,K., Kanki,M., Horinouchi,A., Ono,A., Yamada,H., Ohno,Y. and Urushidani,T. (2014) miRNA expression atlas in male rat. *Sci. Data*, **1**, 140005.
9. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
10. Backes,C., Leidinger,P., Altmann,G., Wuerstle,M., Meder,B., Galata,V., Mueller,S.C., Sickert,D., Stahler,C., Meese,E. *et al.* (2015) Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. *Anal. Chem.*, **87**, 8910–8916.

11. Backes,C., Sedaghat-Hamedani,F., Frese,K., Hart,M., Ludwig,N., Meder,B., Meese,E. and Keller,A. (2016) Bias in high-throughput analysis of miRNAs and implications for biomarker studies. *Anal. Chem.*, **88**, 2088–2095.
12. Meder,B., Backes,C., Haas,J., Leidinger,P., Stahler,C., Grossmann,T., Vogel,B., Frese,K., Giannitsis,E., Katus,H.A. *et al.* (2014) Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin. Chem.*, **60**, 1200–1208.
13. Nagy,C., Maheu,M., Lopez,J.P., Vaillancourt,K., Cruceanu,C., Gross,J.A., Arnovitz,M., Mechawar,N. and Turecki,G. (2015) Effects of postmortem interval on biomolecule integrity in the brain. *J. Neuropathol. Exp. Neurol.*, **74**, 459–469.
14. Lv,Y.H., Ma,K.J., Zhang,H., He,M., Zhang,P., Shen,Y.W., Jiang,N., Ma,D. and Chen,L. (2014) A time course study demonstrating mRNA, microRNA, 18S rRNA, and U6 snRNA changes to estimate PMI in deceased rat's spleen. *J. Forensic Sci.*, **59**, 1286–1294.
15. Ibberson,D., Benes,V., Muckenthaler,M.U. and Castoldi,M. (2009) RNA degradation compromises the reliability of microRNA expression profiling. *BMC Biotechnol.*, **9**, 102.
16. Cloonan,N., Wani,S., Xu,Q., Gu,J., Lea,K., Heater,S., Barbacioru,C., Steptoe,A.L., Martin,H.C., Nourbakhsh,E. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, R126.
17. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
18. Swierniak,M., Wojcicka,A., Czetwertynska,M., Stachlewska,E., Maciag,M., Wiechno,W., Gornicka,B., Bogdanska,M., Koperski,L., de la Chapelle,A. *et al.* (2013) In-depth characterization of the microRNA transcriptome in normal thyroid and papillary thyroid carcinoma. *J. Clin. Endocrinol. Metab.*, **98**, E1401–E1409.
19. Callis,T.E., Chen,J.F. and Wang,D.Z. (2007) MicroRNAs in skeletal and cardiac muscle development. *DNA Cell Biol.*, **26**, 219–225.
20. Thum,T., Catalucci,D. and Bauersachs,J. (2008) MicroRNAs: novel regulators in cardiac development and disease. *Cardiovasc. Res.*, **79**, 562–570.
21. Liu,S., Tetzlaff,M.T., Liu,A., Liegl-Atzwanger,B., Guo,J. and Xu,X. (2012) Loss of microRNA-205 expression is associated with melanoma progression. *Lab. Invest.*, **92**, 1084–1096.
22. Karaayvaz,M., Pal,T., Song,B., Zhang,C., Georgakopoulos,P., Mehmood,S., Burke,S., Shroyer,K. and Ju,J. (2011) Prognostic significance of miR-215 in colon cancer. *Clin. Colorectal Cancer*, **10**, 340–347.
23. Akamatsu,S., Hayes,C.N., Tsuge,M., Miki,D., Akiyama,R., Abe,H., Ochi,H., Hiraga,N., Imamura,M., Takahashi,S. *et al.* (2015) Differences in serum microRNA profiles in hepatitis B and C virus infection. *J. Infect.*, **70**, 273–287.
24. Krauskopf,J., Caiment,F., Claessen,S.M., Johnson,K.J., Warner,R.L., Schomaker,S.J., Burt,D.A., Aubrecht,J. and Kleinjans,J.C. (2015) Application of high-throughput sequencing to circulating microRNAs reveals novel biomarkers for drug-induced liver injury. *Toxicol. Sci.*, **143**, 268–276.
25. Pirola,C.J., Fernandez Gianotti,T., Castano,G.O., Mallardi,P., San Martino,J., Mora Gonzalez Lopez Ledesma,M., Flichman,D., Mirshahi,F., Sanyal,A.J. and Sookoian,S. (2015) Circulating microRNA signature in non-alcoholic fatty liver disease: from serum non-coding RNAs to liver histology and disease pathogenesis. *Gut*, **64**, 800–812.
26. Xu,J., Wu,C., Che,X., Wang,L., Yu,D., Zhang,T., Huang,L., Li,H., Tan,W., Wang,C. *et al.* (2011) Circulating microRNAs, miR-21, miR-122, and miR-223, in patients with hepatocellular carcinoma or chronic hepatitis. *Mol. Carcinog.*, **50**, 136–142.
27. Akat,K.M., Moore-McGriff,D., Morozov,P., Brown,M., Gogakos,T., Correa Da Rosa,J., Mihailovic,A., Sauer,M., Ji,R., Ramarathnam,A. *et al.* (2014) Comparative RNA-sequencing analysis of myocardial and circulating small RNAs in human heart failure and their utility as biomarkers. *Proc. Natl Acad. Sci. USA*, **111**, 11151–11156.
28. Gomes,C.P., Oliveira-Jr,G.P., Madrid,B., Almeida,J.A., Franco,O.L. and Pereira,R.W. (2014) Circulating miR-1, miR-133a, and miR-206 levels are increased after a half-marathon run. *Biomarkers*, **19**, 585–589.
29. Cacchiarelli,D., Legnini,I., Martone,J., Cazzella,V., D'Amico,A., Bertini,E. and Bozzoni,I. (2011) miRNAs as serum biomarkers for Duchenne muscular dystrophy. *EMBO Mol. Med.*, **3**, 258–265.

3.9 Distribution of microRNA biomarker candidates in solid tissues and body fluids

This article is available under: https://doi.org/10.1080/15476286.2016.1234658

This article is available under: https://doi.org/10.1080/15476286.2016.1234658

This article is available under: https://doi.org/10.1080/15476286.2016.1234658

This article is available under: https://doi.org/10.1080/15476286.2016.1234658

This article is available under: https://doi.org/10.1080/15476286.2016.1234658

3.10 Spring is in the air: seasonal profiles indicate vernal change of miRNA activity

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

This article is available under: https://doi.org/10.1080/15476286.2019.1612217

3.11 *The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals*

Published online 2 April 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 9* ***4431–4441***
doi: 10.1093/nar/gkz227

The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals

Tobias Fehlmann[1], Christina Backes[1], Marcello Pirritano[2,3], Thomas Laufer[1,4], Valentina Galata[1], Fabian Kern[1], Mustafa Kahraman[1,4], Gilles Gasparoni[5], Nicole Ludwig[6], Hans-Peter Lenhof[7,8], Henrike A. Gregersen[9], Richard Francke[10], Eckart Meese[6], Martin Simon[2,3] and Andreas Keller[1,8,*]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Molecular Cell Dynamics, Center for Human and Molecular Biology, Saarland University, 66123 Saarbrücken, Germany, [3]Molecular Cell Biology and Microbiology, University of Wuppertal, 42097 Wuppertal, Germany, [4]Hummingbird Diagnostics GmbH, 69120 Heidelberg, Germany, [5]Department of Genetics, Center for Human and Molecular Biology, Saarland University, 66123 Saarbrücken, Germany, [6]Department of Human Genetics, Saarland University Hospital, 66421 Homburg, Germany, [7]Chair for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [8]Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [9]Zoological Garden Neunkirchen, 66538 Neunkirchen, Germany and [10]Zoological Garden Saarbrücken, 66121 Saarbrücken, Germany

Received March 08, 2019; Editorial Decision March 20, 2019; Accepted March 29, 2019

ABSTRACT

The repertoire of small noncoding RNAs (sncRNAs), particularly miRNAs, in animals is considered to be evolutionarily conserved. Studies on sncRNAs are often largely based on homology-based information, relying on genomic sequence similarity and excluding actual expression data. To obtain information on sncRNA expression (including miRNAs, snoRNAs, YRNAs and tRNAs), we performed low-input-volume next-generation sequencing of 500 pg of RNA from 21 animals at two German zoological gardens. Notably, none of the species under investigation were previously annotated in any miRNA reference database. Sequencing was performed on blood cells as they are amongst the most accessible, stable and abundant sources of the different sncRNA classes. We evaluated and compared the composition and nature of sncRNAs across the different species by computational approaches. While the distribution of sncRNAs in the different RNA classes varied significantly, general evolutionary patterns were maintained. In particular, miRNA sequences and expression were found to be even more conserved than previously assumed. To make the results available for other researchers, all data, including expression profiles at the species and family levels, and different tools for viewing, filtering and searching the data are freely available in the online resource ASRA (Animal sncRNA Atlas) at https://www.ccb.uni-saarland.de/asra/.

INTRODUCTION

Since the establishment of the central dogma of molecular biology by Crick (1), for decades the main role of RNAs was believed to be either in the transfer of information between DNA and proteins (mRNAs) or in housekeeping functions (tRNAs, rRNAs). With the discovery of microRNAs in the early 1990s (2), research on small noncoding RNAs (sncRNAs) and later on long noncoding transcripts (3) gained traction. Moreover, advances in high-throughput sequencing technology that allowed the sequencing of millions to billions of small RNA fragments with reasonable effort and cost (4) led to a further growth in the field. Via sequencing-based approaches, the number of identified sncRNAs, especially of miRNAs, increased markedly in just a few years. While the reference repository miRBase (5) was established in the year 2000 with only 222 miRNAs in five species, the most recent version stores 48 885 miRNAs in 271 species. miRCarta (6), a database that collects mature miRNAs independently of the organism, suggests up to 44 347 miRNA candidates; however, only a fraction of these can be assumed to actually be true miRNAs. Because miRNAs have been described in a variety of organisms, their assumed conservation is frequently used to identify additional miRNAs in related species by homology- and sequence-based approaches (7–11), which often exclude expression profiling. Interestingly, the expression patterns of homologous miRNAs also appear to be comparable between organs in dif-

[*]To whom correspondence should be addressed. Tel: +49 681 302 68611; Fax: +49 681 302 58094; Email: andreas.keller@ccb.uni-saarland.de

ferent species, as we successfully showed for human and rat (12).

One of the most commonly performed types of study on sncRNAs is biomarker discovery analysis (13–15). Here, human serum, plasma or blood cells are sequenced, or expression profiling using microarrays or real-time quantitative reverse transcription PCR (RT-qPCR) is performed. Blood cells are especially suitable for this as they contain many hundred to over 1000 human miRNAs (12,16). It has already been demonstrated that the use of standardized protocols for collecting and analysing blood-borne miRNA profiles has huge potential for comparing biomarker profiles across different human pathologies (17,18). Because blood can be obtained in a standardized manner and miRNA expression patterns are technically very stable, it is easy to accurately compare expression between different animal species. In particular, dried blood spots (19) (DBS) or microsampling devices (20) appear to be well suited as containers for miRNAs. While such decentralized collection kits are perfectly suited to collecting samples from different sites, the small amount of RNA that can be purified presents a challenge for further investigations. Previously, analyses based on DBS were mostly limited to microarrays and RT-qPCR, but excluded next-generation sequencing (NGS). However, the application of NGS was mandatory for our study to be able to compare the total sncRNA repertoires amongst different species. Thus, we developed a novel low-input-volume NGS protocol to facilitate sequencing from capillary microsampling devices starting with only 50 pg of RNA (20).

In the present study, we sequenced blood samples of 21 animals collected at two regional German zoos: in Saarbrücken and Neunkirchen. A phylogenetic tree of the animals is presented in Figure 1. The primary data analysis was performed with our tool miRMaster (21). We analysed and compared the read profiles as well as the distribution and composition of small RNAs across species. In addition, an online resource for the collected data was implemented and is freely available at: https://www.ccb.uni-saarland.de/asra/. This resource provides access to all detected sncRNAs, their families and their expression patterns across all species in this study. In summary, the compiled dataset and associated online web server constitute a valuable resource for sncRNA research, either for finding and validating miRNA candidates because of their conservation, or for general research on evolutionary aspects of sncRNAs.

MATERIALS AND METHODS

Sample collection

We collected 21 animal samples from regional zoos in Saarbrücken and Neunkirchen (Germany) comprising 19 different species. In addition, we collected four human samples as a reference. All blood samples were collected with the Mitra™ microsampler device (Neoteryx, CA). The samples were collected from remaining blood samples in the context of veterinary examinations. No additional examinations were performed with the animals. The study was permitted by the regional authority, the State Office for Consumer Protection (Landesamt für Verbraucherschutz). Human blood samples were collected from volunteers with informed consent. An overview of the samples in this study with their corresponding taxonomic classification is given in Table 1. Metadata containing the age, gender, as well as the health condition for each specimen are available in Supplementary Table S1.

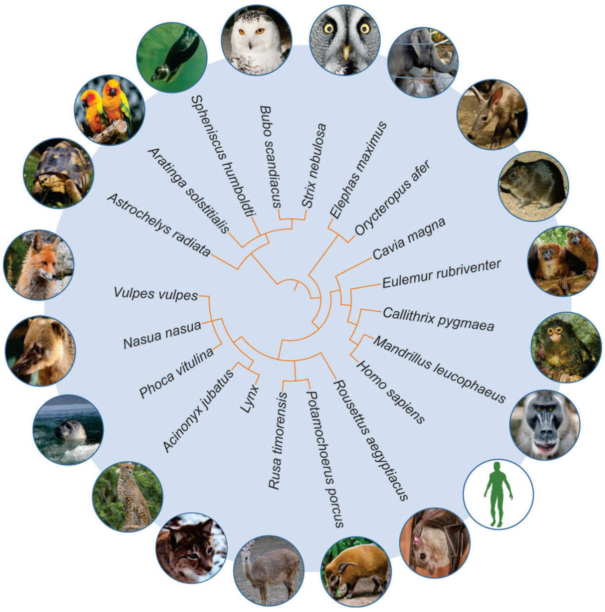


Figure 1. Circular taxonomy tree based on the species that were sequenced in our study.

RNA extraction and sequencing

Animal blood was collected onto Mitra™ collection devices (Neoteryx, CA) and dried at least for 2 h. Small RNAs were extracted by a modified version of the manufacturer's procedure using the miRNeasy Serum/Plasma Kit (Qiagen, Hilden, Germany). Size distribution and concentration were analysed using Agilent Bioanalyzer small RNA chips (Agilent Technologies, Santa Clara, CA). A total of 500 pg of sRNA with a size range of ∼15–150 nt was subjected to library preparation using a ligation-free procedure involving 3'-polyadenylation and template switch-based cDNA synthesis using the CATS sRNA-seq Kit (Diagenode, Liege, Belgium), omitting any dephosphorylation to enrich 3'-OH. Library size enrichment was carried out using 1.8 vol AMPure XP beads (Beckman Coulter, Krefeld, Germany) to achieve the enrichment of libraries containing RNAs larger than 15–20 nt (library size >160 bp). Libraries were multiplex-sequenced in an Illumina HiSeq 2500 platform in high-output mode with 50 cycles, except for common seal (1), human (3), pygmy marmoset, radiated tortoise and red-bellied lemur that were (re)sequenced with 40 cycles. Lynx (2) was sequenced with 47 cycles.

Table 1. Overview of the sequenced species ordered by phylogeny, their taxonomic classification, their total generated reads and remaining valid reads after filtering and trimming, as well as the availability of a genome assembly

Taxid	Species	Superorder	Order	Total reads (Mio)	Valid reads (Mio)	Genome
9568	*Mandrillus leucophaeus*	*Euarchontoglires*	*Primates*	72.65	52.19	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	25.45	12.14	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	15.46	10.02	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	16.98	9.87	✓
9606	*Homo sapiens*	*Euarchontoglires*	*Primates*	24.50	19.26	✓
9493	*Callithrix pygmaea*	*Euarchontoglires*	*Primates*	38.80	27.76	✗
34829	*Eulemur rubriventer*	*Euarchontoglires*	*Primates*	35.50	21.09	✗
297387	*Cavia magna*	*Euarchontoglires*	*Rodentia*	36.54	25.38	✗
273791	*Potamochoerus porcus*	*Laurasiatheria*	*Artiodactyla*	32.85	24.68	✗
1088130	*Rusa timorensis*	*Laurasiatheria*	*Artiodactyla*	37.23	25.41	✗
9720	*Phoca vitulina*	*Laurasiatheria*	*Carnivora*	24.57	16.15	✗
9720	*Phoca vitulina*	*Laurasiatheria*	*Carnivora*	23.73	14.95	✗
9651	*Nasua nasua*	*Laurasiatheria*	*Carnivora*	46.87	34.31	✗
9627	*Vulpes vulpes*	*Laurasiatheria*	*Carnivora*	29.23	20.26	✓
13124	*Lynx*	*Laurasiatheria*	*Carnivora*	47.84	22.31	✗
13124	*Lynx*	*Laurasiatheria*	*Carnivora*	28.72	17.23	✗
32536	*Acinonyx jubatus*	*Laurasiatheria*	*Carnivora*	30.62	20.65	✓
9407	*Rousettus aegyptiacus*	*Laurasiatheria*	*Chiroptera*	33.75	24.99	✓
9783	*Elephas maximus*	*Afrotheria*	*Proboscidea*	97.67	63.16	✗
9818	*Orycteropus afer*	*Afrotheria*	*Tubulidentata*	36.68	26.45	✓
371907	*Bubo scandiacus*	*Neognathae*	*Strigiformes*	58.79	38.58	✗
126836	*Strix nebulosa*	*Neognathae*	*Strigiformes*	37.81	27.92	✗
176015	*Aratinga solstitialis*	*Neognathae*	*Psittaciformes*	43.77	28.29	✗
9240	*Spheniscus humboldti*	*Neognathae*	*Sphenisciformes*	72.75	53.41	✗
66190	*Astrochelys radiata*	*Chelonia*	*Testudines*	25.24	17.76	✗

Bioinformatics

Sample preprocessing. All samples were trimmed and cleaned using miRMaster (21). In detail, we first removed the template switch motif, i.e. the first three bases of the reads. Then, we removed the bases resulting from the polyadenylation process. Therefore, we first checked the reads for adenine homopolymers with at least 13 bases and at most one mismatch and, if no match was found, we relaxed the requirement for an adenine homopolymer with at least five bases and no mismatch starting at position 15 of the read. Finally, we removed sequencing adapter contamination. The quality filtering was performed using default parameters together with a sliding window of four bases and a quality threshold of 20. The resulting reads that were shorter than 17 nt were discarded.

Statistics and visualizations. All statistical tests were computed using the free statistical programming language R (22) (version 3.4.4). If not specified otherwise, reported *P*-values were adjusted for multiple testing using the Benjamini-Hochberg procedure (23). Cramer's V was computed using the R package rcompanion (24). Wilcoxon-rank sum test was applied when the data did not follow normal distribution according to Shapiro–Wilk test. Plots were generated using the R packages ggplot2 3.1.0 (25) and pheatmap 1.0.12.

Sample distance estimation and similarity to NCBI phylogenetic tree. We computed Mash sketches for all samples (using Mash 2.0 (26)) with a *k*-mer size of 17 and a signature size of 1000 and used them to estimate the pairwise sample distances. Reads were subsampled using Seqtk 1.2. We constructed a phylogenetic tree using the neighbour-joining approach (27) implemented in the R-package phangorn (28) and visualized it using the Interactive Tree of Life (29). The similarity to the phylogenetic tree provided by NCBI was computed using the normalized Robinson-Founds distance. To be able to compare both trees, we collapsed the nodes of the same species. We determined the significance of the similarity of both trees by creating 100 000 random trees with 20 leaves, labeled by the analysed species and comparing them with the NCBI tree. We then tested if the resulting distances were smaller than the computed distance and derived from this the *P*-value.

Rfam. We downloaded all Rfam family sequences from the Rfam FTP server (ftp://ftp.ebi.ac.uk/pub/databases/Rfam, version 13, accessed on 27/3/2018). Then, we determined that sequences were related to Metazoa by performing an SQL query against the Rfam database, and selected them accordingly. To this end, we used the following SQL query:

```
SELECT fr.rfam_acc, fr.rfamseq_acc,
   fr.seq_start, fr.seq_end, f.type
FROM full_region fr, rfamseq rf,
   taxonomy tx, family f
WHERE rf.ncbi_id = tx.ncbi_id
AND f.rfam_acc = fr.rfam_acc
AND fr.rfamseq_acc = rf.rfamseq_acc
AND tx.tax_string LIKE '
AND is_significant = 1
```

Next, we mapped all samples against the Metazoa Rfam sequences using RazerS 3 (30), while requiring at least 95% identity and allowing only forward mappings. We determined the RNA composition based on the RNA class annotations of each family. If a read mapped to multiple classes, it was counted in full for each.

miRNA homology determination. We collected the miRNA sequences of miRBase v22, miRCarta v1.0 and MirGeneDB 2.0 via their respective websites (accessed on 18 July 2018). To determine the expression of each miRNA, we mapped the samples against the databases with Bowtie (31) (version 1.1.2), while allowing no mismatches and disabling mapping against the reverse complement, using the following command:

```
bowtie -f -v 0 -a --fullref --norc
   -S <reference_mirnas_idx> <sample.fa>
```

To ensure that each read corresponds to a real miRNA, we discarded all reads with lengths different from those of their mapped miRNA. A miRNA was considered to be expressed in a species if it was present in at least one of its samples.

miRNA expression and potential precursor determination. MiRNAs found in any of the three considered databases were first clustered according to 90% sequence similarity using vsearch 2.7.1 (32), thereby merging potential isoforms into one cluster. The RPM normalized counts for each cluster were determined by summing up the expression of each miRNA contained in it. MiRNA arms were determined according to their annotation in the databases. Potential precursors were determined for the miRNAs by considering all combinations of 5′ and 3′ miRNAs of precursors of the same precursor family for MirGeneDB, with the same base name for miRBase and according to the exact annotations in miRCarta. MiRNAs that could not be assigned unambiguously to one arm were discarded. Using the thereby obtained potential precursors, we could then compute arm ratio differences to investigate arm switches.

MiRNA candidate prediction. MiRNA candidates were predicted using mirnovo (33) (downloaded on 20 July 2018) with the default parameters, except for the brown-nosed coati, for which we had to increase the required minimum number of isoform variants from 1 to 3 because the program was not terminating with lower numbers. Predicted miRNAs were filtered in a first step by only keeping those that did not map with at least 90% identity to any known miRNA. The mapping was performed with RazerS 3 (version 3.5.8). Subsequently, we built a scoring scheme similar to our tool novoMiRank (34). In a first step, we determined the values of the features used by mirnovo for known miRNAs in our dataset. To this end, we restricted the known miRNAs to those contained in the high-confidence set of miRBase v22, as we recently showed that this subset contains by far the largest fraction of true miRNAs (35). The features of mirnovo do depend not only on the miRNAs but also on the samples. It is thus possible that some miRNAs that are more expressed than others bias the feature distribution. To avoid this bias, we took the mean feature values for every miRNA. We then normalized all features to a mean of zero and a variance of once, since they were all on different scales and computed z-scores for all known miRNAs. To avoid too large influences of single features, we restricted the absolute values to 3. We then computed for every predicted miRNA its distance to the distribution of known miRNAs, for every feature, and reported the mean z-score. As filtering threshold we chose the 80th percentile of the z-scores of known miRNAs, corresponding to 0.8 standard deviations above or below the mean of the known miRNAs.

ASRA. In the web resource, we provide a species specificity index (SSI) for miRNAs and for Rfam families that describe the variability of their expression patterns. It is computed analogously to the tissue specificity index used in our miRNA tissue atlas (12). It allows measurement of the specificity of expression of an miRNA/Rfam family over different species. The SSI ranges from 0 to 1, where values closer to 1 represent molecules expressed in a few or only one species (species-specific molecules) and values closer to 0 represent molecules similarly expressed in many species (well-conserved molecules). To this end, the SSI for an miRNA/Rfam family j is calculated as follows:

$$ssi_j = \frac{\sum_{i=1}^{N}(1 - x_{j,i})}{N - 1}$$

where N corresponds to the total number of species and $x_{j,\,i}$ is the RPM expression of the miRNA/Rfam family j in species i normalized by the maximal expression in any species of miRNA/Rfam family j.

RESULTS

Using the Mitra™ system, we collected a total of 21 specimens from two regional zoos, including 19 animal species, as well as four human samples. The species in this study belong to five different superorders and 11 different orders. The samples were sequenced on an Illumina HiSeq 2500, yielding a total of 973 994 362 reads. After quality filtering and adapter trimming 654 217 441 reads remained and were used for downstream analysis. An overview of the collected samples, their taxonomy and read counts is presented in Table 1. Due to the fact that for only five of the sequenced animal species a genome assembly is available to date, of which all are on scaffold level, no genome mappings were computed. Also, no miRNAs were annotated in any of the considered reference databases. All downstream analyses were performed only with the valid reads.

Read profiles resemble phylogenetic descriptors

One of the core hypotheses in this study is that the differences in read profiles between the species also mirror their known taxonomic classification. To test this hypothesis, we conducted a minHash analysis using Mash (26). The top panel of Figure 2 shows the resulting 2D embedding based on the computed sample Mash distances for superorders (2 A) and orders (2 B). For the superorders, we observe a cluster pattern matching what one would expect from their taxonomy, with the exception of *Afrotheria*. In the more detailed 2D embedding for orders, we see that samples belonging to *Primates*, *Carnivora* and *Strigiformes* cluster together well. Since the amount of reads for our samples varied greatly we wanted to estimate this influence. Therefore, we generated embeddings based on 15 times subsampling of the depth of the smallest sample, for each sample. This way, we ensure that all samples have the same size, while still keeping a realistic sequencing depth. The resulting plots

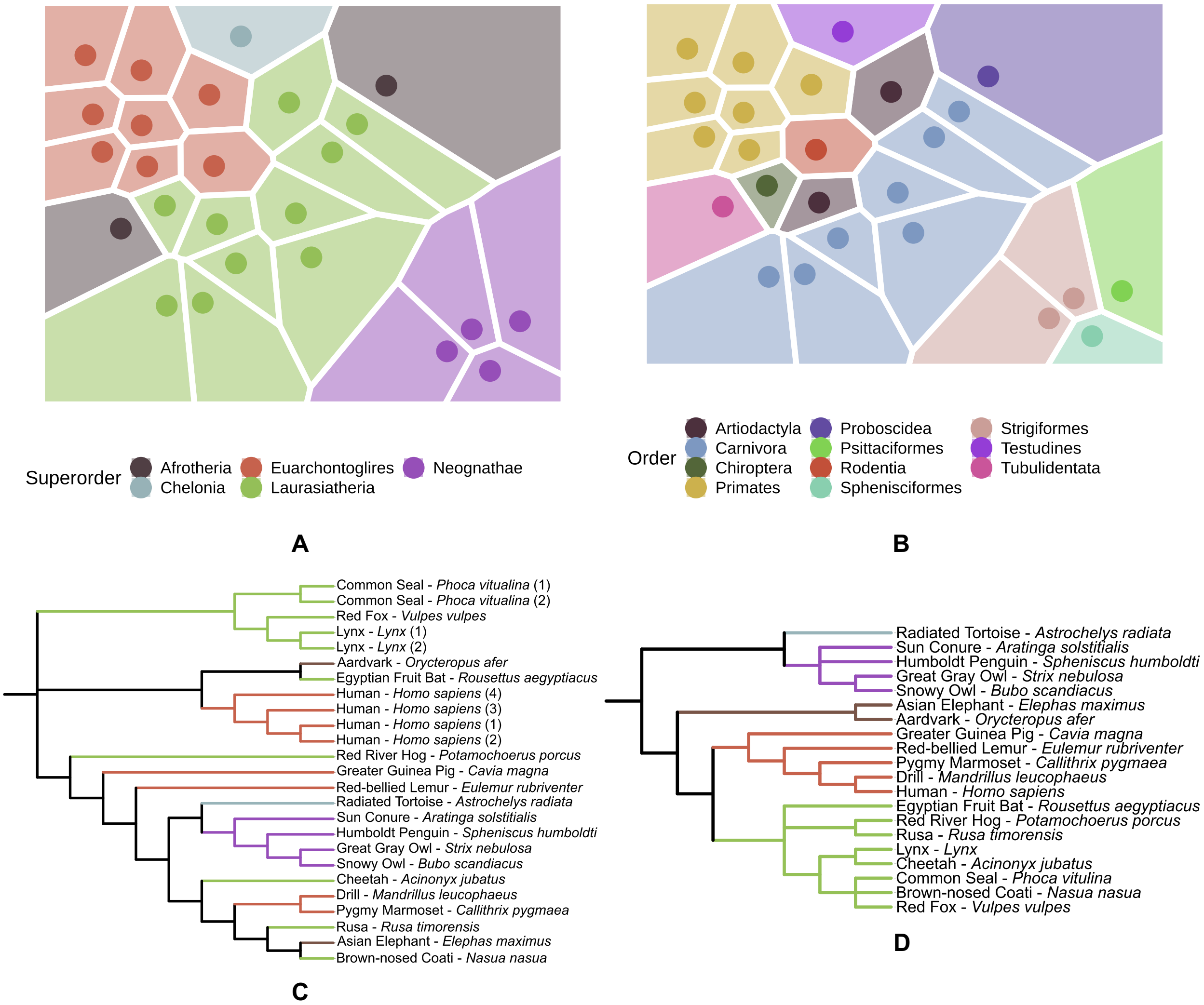


Figure 2. 2D embedding including a Voronoi diagram of the pairwise sample Mash distances for superorders (**A**) and orders (**B**). Each point in the plot represents a sample. Taxonomic tree built using the computed Mash distances of the read profiles at the species level (**C**) in comparison to the taxonomic tree derived from NCBI (**D**). The branches are colored according to the superorder of the corresponding species.

(Supplementary Figure S1) show that the sample depth has only a minor influence on the clustering. To increase the resolution to the species level, we visualized the computed Mash distances as a phylogenetic tree, as shown in the lower panel of Figure 2, in comparison to the phylogenetic tree from NCBI. The biological replicates for human, common seal and lynx cluster together, confirming the reproducibility of the sample collection and sequencing process. For some species, the clustering in the Mash tree matches very well with the partitioning in the NCBI taxonomy tree; for example, the two owls cluster with the Humboldt penguin and the sun conure, which form a larger cluster with the radiated tortoise. Drill and pygmy marmoset also cluster together in both trees; however, the human samples do not cluster with these species as we would expect from the NCBI phylogenetic tree, which is partly related to the heuristic nature of the neighbour-joining algorithm used to create the tree. To quantify the resemblance of both trees, we computed the normalized Robinson-Foulds distance between both trees ($D = 0.8$) and found that it was significantly lower than expected by chance ($P = 4 \times 10^{-5}$). While some of the remaining sample clusters do not fit the known taxonomy perfectly, we still see that, based on the distance of read profiles alone, we can derive evolutionary relationships to a certain extent.

Distribution of sncRNAs varies across species

To obtain an overview of the distribution and composition of sncRNAs across species, we mapped their reads to the sequences from the Rfam database (36) with a threshold of 95% identity. We then evaluated the quality of the mappings

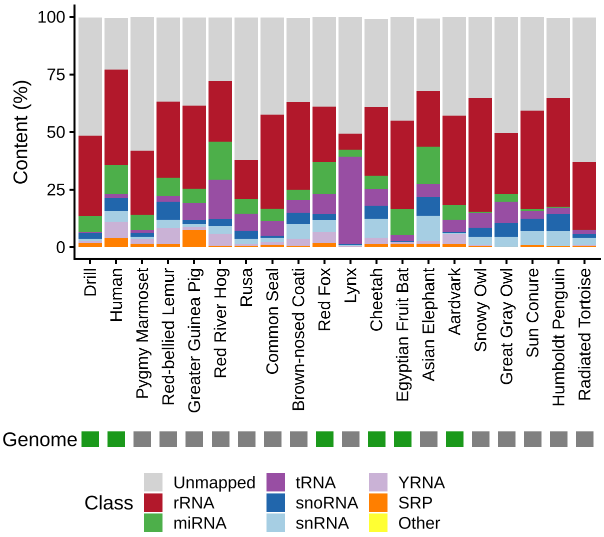


Figure 3. Overview of reads mapped to the different Rfam classes for all species in this study. The colors are ordered according to the median mapping ratio of each class. Classes with mapped reads <0.05% are summarized in the category 'Other'.

by inspecting the distribution of their read lengths after trimming (Supplementary Figure S2) and comparing them with the distribution of the mappings against every RNA class of Rfam (Supplementary Figures S3–10). We observe in all sample peaks at the length of the sequenced reads (minus 3 nt of the template switch motif), i.e. at 47 nt and for some that ran with less cycles at 37 nt. In general, we would expect that for RNA classes that are longer than the read lengths, and which have no known functional fragments, mostly untrimmed reads map. This is the case for rRNAs where we observe mainly untrimmed reads. It holds also for snRNAs, where only in few species over 15% of the reads shorter than 30 nt map. Reads mapping to SRP RNAs are mainly untrimmed reads as well; however, in some species the length of the mapped reads is nearly evenly distributed. YRNAs, as well as tRNAs, are either mostly covered by untrimmed reads or reads in the length of YRNA and tRNA fragments (around 26 nt and around 32 nt). For reads mapping to miRNAs, we observe clear mapping patterns that show peaks at 21–22 nt, with mostly no mapping read exceed a length of 24 nt. Considering snoRNAs, we observe mostly mappings of untrimmed reads, except for some species with peaks around 26 nt. Finally, all other mapping reads are composed mostly of untrimmed reads or short reads around 20 nt. The overall results of the mapping distribution are presented per sample in Supplementary Figure S11 and summarized per species by taking the average mapping fraction in Figure 3. As expected, in almost all species, the most dominant read fraction is represented by rRNAs. However, the percentages vary substantially across species: from 7% in lynx to 49.3% in snowy owl, with a median of 35.2%. In particular, the composition of the RNA classes in both lynx samples diverge the most from those in the other species. Here, not only is the rRNA fraction very small, but also the tRNA fraction (which is in median the third most abundant class) represents 38.1% of the sncRNA reads. In most other species, the fraction of tRNAs is under 10% (median 5.5%). The distribution of miRNAs, which are the second most abundant RNA class, also varies amongst the different species, ranging from 0.2% in radiated tortoise to 16.4% in red river hog. Similar patterns could be observed for all other RNA classes. Interestingly, the fraction of miRNAs, but also of YRNAs, was highly underrepresented in all species of the *Neognathae* and *Chelonia* superorder (miRNA mean: 1.1% versus 8.7%, Wilcoxon rank-sum test $P = 5 \times 10^{-6}$; YRNA mean: 0.27% versus 2.9%, Wilcoxon rank-sum test $P = 4 \times 10^{-4}$). The differences in the compositions of RNA classes might also be influenced by the number of unmapped reads. Human reads are much better recovered in Rfam than reads of rusa and radiated tortoise, for example (unmapped: ∼23% versus ∼62%, respectively). We investigated if the mapping rates were associated with the presence of a genome assembly; however, no significant association was found (Wilcoxon rank-sum test (two-sided) $P = 0.968$). A chi-square test of homogeneity showed that all pairwise sample comparisons differ significantly ($P = 0$). Since the *P*-values are strongly affected by large read counts, we also computed the effect sizes using Cramer's V, see Supplementary Table S2. Thereby, we found that the values for samples of the same species (median: 0.16) were significantly smaller (i.e. the class distributions were more similar to each other) than for samples between different species (median: 0.31, Wilcoxon rank-sum test (one-sided) $P = 9 \times 10^{-6}$), highlighting that even though all RNA class distributions were significantly different, the heterogeneity between samples of different species was higher than between samples of the same. To assess if the observed class distributions of some RNA classes are related to each other, we computed all pairwise Spearman correlation coefficients (Supplementary Figure S12) on the number of reads mapped to each class. This showed that miRNA and YRNA levels, as well as snoRNAs and snRNAs, are significantly and positively correlated to each other ($\rho = 0.72$, $P = 6 \times 10^{-4}$ for miRNAs and YRNAs, and $\rho = 0.89$, $P = 3 \times 10^{-5}$ for snoRNAs and snRNAs).

Zoo animals express common miRNA families that are more conserved than previously assumed

We also evaluated the coverage of known miRNA sequences and miRNA families in the different species. To obtain a comprehensive overview, we made use of three different miRNA databases with different scope: miRBase v22 (5), miRCarta v1.0 (6) and MirGeneDB 2.0 (37). miRBase is the gold standard resource for miRNAs; miRCarta also collects many miRNA candidates, of which only a fraction might be true miRNAs; and MirGeneDB collects miRNA genes that are manually curated and validated. We mapped the reads of the different species against the mature miRNA sequences of the three different databases, allowing only exact matches, which means that we count only reads that have exactly the same sequence and length as the sequence deposited in the corresponding database. Figure 4 summarizes the findings for the three databases separately, as well as the results overlapping amongst them. As a me-

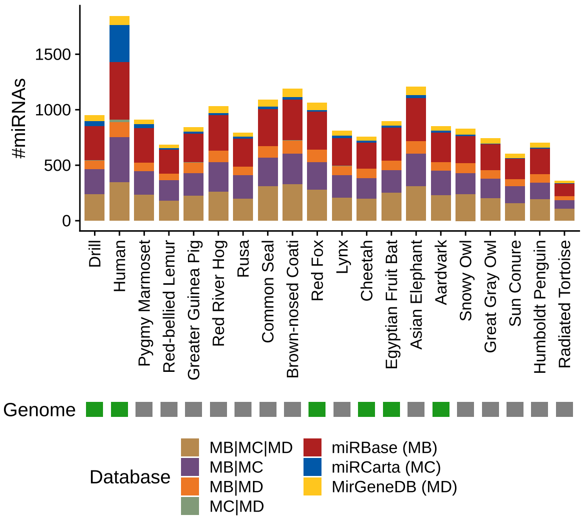


Figure 4. Comparison of mapping the reads of the different species against the three miRNA databases: miRBase, miRCarta and MirGeneDB. The mapping was performed with perfect matches, allowing no mismatches or differences in lengths between read and database sequence. The stacked barplot shows the number of miRNAs found uniquely in the corresponding databases, as well as the different overlaps amongst the databases.

dian, we recovered 847 miRNAs per sample. Because human is the organism with the most annotated miRNAs, we recovered the most miRNA sequences in human (n = 1846), followed by Asian elephant (n = 1210) and brown-nosed coati (n = 1187). At the lower end, the reads of the radiated tortoise sample recovered only 358 miRNAs. We could expect the number of recovered miRNAs to be significantly higher in species with known genome; however, this was not the case (Wilcoxon rank-sum test (one-sided) P = 0.1037). Although a large proportion of the miRNA sequences overlap with references in the three databases or in any combination thereof, we still found many unique hits of the reads, especially for miRNAs from miRBase. While this is surprising at first glance, it can be explained by the difference in set-up between miRCarta and miRBase. In these databases, similar miRNAs are merged into one representative, but miRBase might contain variants of the same miRNA sequence with different lengths. Nonetheless, for assessing which miRNAs actually exist, these sequences uniquely recovered in the different databases might provide new insights, because they appear to be expressed in different species in our study. To this end, we analysed the uniquely recovered sequences in miRBase in more detail. In total, we discovered 862 unique miRBase sequences, of which 44 were found in all 20 species in our deep sequencing approach. Interestingly, most of these have been described in only three different organisms in miRBase, on average. Amongst those 44 recovered sequences, there are many representatives of well-known families, such as let-7, mir-17, mir-103, mir-24, mir-181 and mir-92. Our findings indicate that these miRNAs are expressed in substantially more species than previously assumed and provide new insights into their conservation. If we look at the unique miRBase sequences recovered that have the most miRBase organisms' annotations, but are found in only a few of the species in our analysis, we might conclude that these are either not as evolutionarily conserved or predominantly expressed as isoforms with different sequence lengths, or might even represent artefacts that have been derived by sequence-based homology but not by expression analysis. One such example is the sequence 5′-CUGCCCUGGCCCGAGGGACCGA-3′, which is only found in one species amongst our samples, but is annotated in 10 miRBase organisms. However, if we remove one base at the 3′ end from this, we also find this sequence in seven further organisms in our study and in two from miRBase. Essentially, this shows that this sequence might be a conserved miRNA, but occurs in at least two isoforms of different lengths. The uniquely recovered miRBase sequences, the number of species they cover in our study and in how many miRBase organisms the sequences are annotated are shown in Supplementary Table S3.

Some sncRNAs are processed depending on the superorder of their species

Small noncoding RNAs and especially miRNAs are known to be expressed differently in organisms depending on various factors such as diseases, developmental stages or tissues. Therefore, we asked if we could find such relationships between our species as well, and in particular if this would be related to phylogeny. In a first step, to avoid biases related to isoforms, we clustered all detected miRNAs with an identity of at least 90% together and summed their expression values. Next, we clustered the miRNAs that represented at least 0.1% of the total miRNA expression in the corresponding species and that were present in at least 5 species (see Supplementary Figure S13). There, we observed that the strongest split between the species happened between those of the superorders of Neognathae and Chelonia in comparison to the other three. This is in concordance with our observations made in the previous analyses, as well as with the phylogenetic tree provided by NCBI. One example of miRNA expressed nearly exclusively in Neognathae and Chelonia is miR-2188-5p. This miRNA is expressed with a median of over 30 000 reads in those species, whereas in others we found it in at most 328 reads. In opposition, for example miR-423-3p is mostly expressed in Afrotheria, Euarchontoglires and Laurasiatheria (median of over 25 000 reads) but nearly not in Chelonia and Neognathae (at most 467 reads). We also evaluated if either 5′or 3′ miRNAs were over-represented amongst the evaluated miRNAs; however, their numbers were very similar (66 5′ miRNAs, 63 3′ miRNAs and 48 either undetermined or miRNAs that have been annotated on 5′ and 3′ positions). The observed differences led us to the question if there were potential miRNA precursors that indicated arm switches between species of different superorders. Supplementary Figure S14 shows the fraction of 5′ minus 3′ miRNA reads (1 being thus precursors exclusively expressing their 5′ miRNA and -1 their 3′ miRNA) of potential precursors, derived from the known annotations. We see that most precursors express mainly one form across all species. However, there are some for which there is no clear form. We decided to investigate those further, in particular regarding differences at the superorder level and found nine potential precursors with large differences between the Neognathae and Chelonia superorders in contrast to the Afrotheria, Euarchontoglires and Laurasiatheria superorders (see Supplementary Figure S15). However, differential processing seems to be not only limited to miRNAs, since we found for example different processing profiles for fragments of SNORD14 enriched in most species at the 5′ end, but showing clear preferences for fragments at the 3′ end in great gray owl, red fox and sun conure, as shown in Supplementary Figure S16.

Gender and health condition have limited impact in cross-species RNA expression

Others and we have shown that expression levels of certain sncRNAs, in particular miRNAs, are driven by gender or disease conditions (38–40). Therefore, we evaluated if we could observe different expression levels of Rfam families or miRNAs according to the gender or health conditions (unaffected versus affected) of our sequenced species. We did not perform a more fine grained comparison by disease, since the group sizes would have been too small and some miRNAs, such as miR-144-5p, have been shown to be deregulated independent of the disease in human (39). While significantly differing miRNA and Rfam family levels were found according to a two-sided Wilcoxon rank-sum test (gender specific: RF01412 ($P = 0.013$), miR-224 ($P = 0.026$); health condition specific: RF00009 ($P = 0.0025$), miR-238|miR-548c|miR-1842 ($P = 0.009$)), none remained significant after adjustment for multiple testing. Therefore, we conclude that the impact of these variables in a cross-species setup is too small and that differences between the species dominate the expression levels.

Many miRNA candidates are not covered by known databases

In addition to known miRNAs from the databases above, it is likely that there are other small noncoding RNAs that have not yet been annotated. A mapping-based analysis using a reference genome usually supports the discovery of these candidates. Because, for the majority of the animals included in this study, no reference genome is available, we applied mirnovo for genome-free miRNA prediction (33). First, we assessed how many known miRNAs can be recovered by a run of this tool. Figure 5A shows a stacked barplot for the number of recovered miRNAs deposited in the databases miRBase, miRCarta and MirGeneDB. In this case, we defined a positive hit if the reads mapped with at least 90% identity to the miRNA sequence in a database taking into account mismatches and differences in length. The prediction algorithm recovers most known miRNAs for human, followed by lynx, Egyptian fruit bat and common seal. In contrast to the comparison of the perfect matches above, we see that the largest fraction of recovered miRNAs is shared by all three databases for each organism and that miRCarta entries contribute the largest proportion. Still, the number of recovered miRNAs is moderate overall; even for human, we recover only 360 miRNAs. As a median, we recover only 40.5 miRNAs across all samples. Second, we analysed the results of the mirnovo algorithm by excluding known miRNAs and illustrate the numbers of novel predictions in Supplementary Figure S17. Here, as a median, approximately 575 miRNAs per species remain. The organism yielding the most candidates is sun conure, with more than 2000 predicted miRNAs, followed by Asian elephant with 1298. Because the gap between known recov-

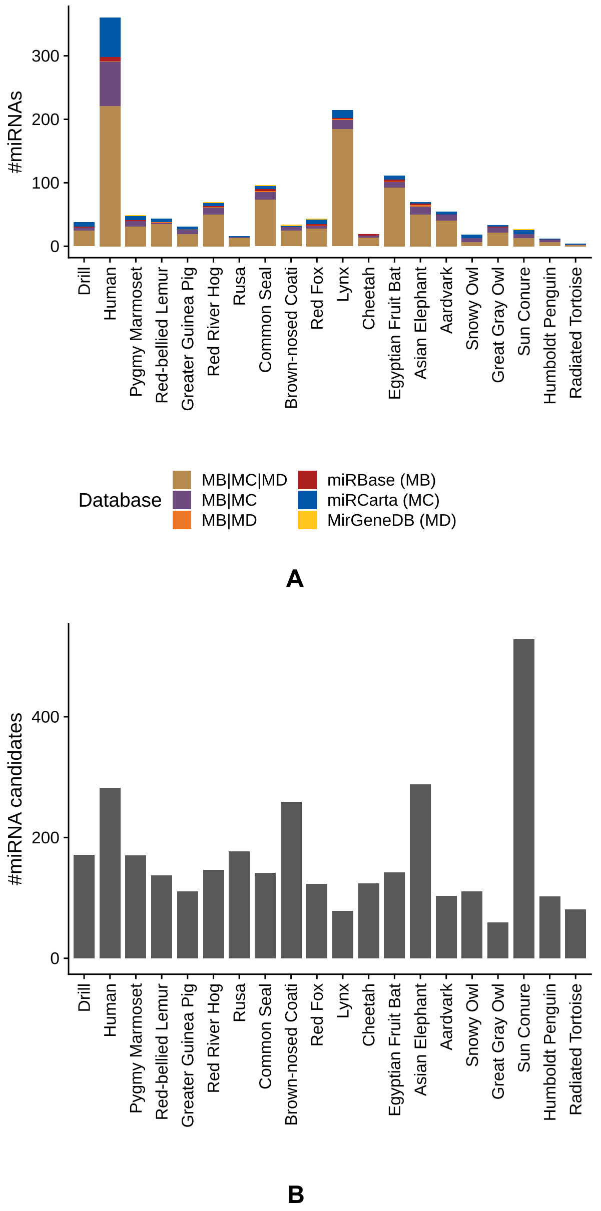


Figure 5. Prediction of novel miRNAs with the tool mirnovo. (**A**) Comparison of recovered known miRNAs deposited in the three databases: miRBase, miRCarta and MirGeneDB. For the mapping, we required at least 90% identity between read and database sequence. The stacked barplot shows the number of miRNAs found uniquely in the corresponding databases, as well as the different overlaps amongst the databases. (**B**) Number of novel miRNAs predicted by mirnovo and filtered by us for the samples in this study.

ered miRNAs and novel miRNAs is quite large, it is questionable how many of the predicted candidates represent true positive findings. To increase the likelihood of predicting true miRNAs, we applied a score filtering similar to novoMiRank (34), based on the features of mirnovo. The obtained scores (see Supplementary Figure S18) highlight that many predicted miRNAs are very different from the miRNAs of the high confidence set of miRBase. By filtering the predictions according to their scores, we reduced the number of predictions by 4-fold in median, as show in Figure 5B, while the number of recovered miRNAs dropped in median only by 2-fold (see Supplementary Figure S19). The results of the filtered mirnovo analysis are available in our online repository.

ASRA: the online resource

In the previous sections, we provide only a snapshot of the potential analyses that are possible using the NGS dataset, excluding many further considerations, such as animal-specific miRNA arm expression preferences, isoforms and others. To make our findings and data easily accessible to others and to promote secondary analyses, we implemented the online resource ASRA (Animal sncRNA Atlas), available at https://www.ccb.uni-saarland.de/asra/. ASRA consists of five major modules. First, we provide an overview of all studied samples and display their read profile similarity in comparison to their phylogenetic annotations, represented as a 2D embedding plot and a phylogenetic tree. Second, users can search specific miRNAs or Rfam families in the databases considered here (miRBase, miRCarta, MirGeneDB and Rfam) and display their expression in all species (for an example, see Supplementary Figure S20). Thereby, the total read counts or expression normalized as the reads per million (RPM) can be shown, as well as the expression of known similar miRNAs (known miRNAs with 90% similarity to the selected one). In addition, a species specificity index is shown for each entry, which indicates whether the displayed RNA is preferentially expressed in few species (values closer to 1) or ubiquitously in all species (values closer to 0). Third, each organism and considered database can be browsed separately; for example, for each organism we provide an overview of the number of reads and their mapped fraction, as well as their class distribution. In addition, detailed mapping information, such as total reads and average RPM, are displayed for the three analysed miRNA databases, the predicted miRNA candidates, the Rfam RNA families as well as their Gene Ontology terms. In particular, for Rfam RNA families, we provide coverage plots with the average RPM at each position of the 500 most expressed family members. All tables can be filtered according to their miRNA/RFAM IDs, their expression or the number of samples in which the sequence was found. Because Rfam families are composed of many sequences, we provide a detailed view for each family and species, which comprises the fourth usability feature. Users can then see if the detected parts of the family are common to many family members or if they are specific to few members. Furthermore, we enable the family coverage profiles to be directly compared amongst different species, which can highlight differences such as miRNA arm expression preferences (arm switches). Finally, users can search nucleotide sequences, either exactly or as part of a read, in all samples of the database and inspect their distribution amongst all species.

DISCUSSION

High-throughput sequencing in combination with microsampling devices allows the generation of data from species for which normal sample collection would be challenging. In our study, we collected blood from a variety of different species at German zoos and compared their small noncoding RNA profiles.

In the first steps of data analysis, quality filtering removed a considerable number of reads. This is probably due to two factors: as we used a minimally invasive method for sampling peripheral blood, the amount of RNA was indeed limited. We consequently chose a library preparation protocol suitable for low input amounts based on ligation-free template-switching cDNA generation. To this end, we used total small RNA fractions from precipitation-free isolation from dried blood without further size exclusion. As such, a high number of very small reads (shorter than 17 nt) were obtained and thus discarded. Next, we used 3′ polyadenylation of small RNAs before reverse transcription, which then requires the trimming of poly(A) stretches. Here, any small RNA with a poly(A) region is trimmed, as we cannot differentiate this from *in vitro* poly(A). For the unmapped fraction of reads and also for species for which, to date, no genome is available, it is unlikely that we sequenced many RNA degradation products, as we omitted any dephosphorylation and therefore enriched the library for 3′-OH RNAs.

Analysing the similarity of the read profiles by computing the Mash distances revealed that most of the samples of the same superorders and orders clustered together. Even at the species level, we still found two groups (birds and primates) that clustered in a way that was comparable to the phylogenetic taxonomy in NCBI. To the best of our knowledge, this is the first study showing that *k*-mer profiles derived from small RNA reads across many species still maintain the known evolutionary relationships.

Upon considering the distribution of RNA classes across species, we could not observe a clear pattern. As expected, rRNA constituted the dominant fraction in most species, with some exceptions. The number of reads that could be mapped to the Rfam classes varied enormously amongst the species. Human had the best coverage, but is also amongst the best annotated and most researched organisms. Relating these distributions to the differing amount of annotations known for many species, it seems reasonable that RNA classes are distributed heterogeneously. However, this is certainly also related to the fact that some organisms are more in the research focus than others. As the other animals in our study are not model organisms, it is possible that their unmapped reads belong to RNA families that have not yet been annotated in Rfam or otherwise present sequencing artefacts. Astonishingly, we found that the tRNA fraction was incredibly high in both lynx samples. As we found similar extreme distributions for both samples, this reduces the likelihood of sequencing or library preparation errors.

Therefore, we hypothesize that this could be related to the physiological or even pathophysiological condition of the Lynx that has not been diagnosed so far, especially since tRNA overexpression has often been associated with various cancer types in human (41–43). Interestingly, we found that miRNAs and YRNA levels were positively correlated, suggesting that even though their biogenesis pathways are different (44) they might share, potentially complementary, functions. We also found that the levels of snoRNAs and snRNAs correlated positively, which is not surprising, as they both belong to the upper class of small nuclear RNAs that guide RNA processing proteins.

The evaluation of the expression of sncRNAs in the context of their phylogeny highlighted that large differences that can be observed between some superorders, and in particular between Neognathae and Chelonia in comparison to the others of this study. We even found examples of potential precursors that showed preferential arm expressions depending on their superorders. Nevertheless, these findings are of course limited by the size of groups, and more samples would be needed for higher confidence. In particular, arm expression comparisons can be difficult, due to the fact that precursors containing the same or similar miRNAs do not necessarily exist in all species. Further evidence, in particular via genome assemblies, would help to reduce this limitation.

The recovery of deposited miRNA sequences from three miRNA databases highlighted that miRBase contains the highest number of unique sequences, but also include numerous redundant variations of sequences belonging to the same family. We showed that known miRNAs are available in more species than previously assumed and other ones might be expressed predominantly as different isoforms.

For the prediction of novel miRNAs from NGS data, we chose mirnovo (33) because this tool does not require a reference genome. To obtain an estimate of how well this prediction works, we counted how many known miRNA sequences can be recovered with the prediction. Although we used a very lenient mapping strategy, a median of only about 40.5 miRNAs were found per organism. In contrast, the tool predicted more than 10 times as many novel candidates per organism. By applying a filtering approach and thus reducing the predictions by 4-fold, we expect to have increased the ratio of true positives considerably. Because we cannot verify these results experimentally, it remains unclear how many true positive findings the predictions actually contain.

While our study describes expression patterns of sncRNAs in blood cells for a large collection of animals and provides fascinating new insights into the distribution and conservation of sncRNAs, certain limitations of the present study need to be considered and discussed. First, the samples were collected during veterinary examinations, including routine examinations but also blood collection of animals with pathologies. These factors might be reflected in the patterns of sncRNAs, but according to our experience from human samples, such effects are rather moderate compared with the variations that we observe here. A more important factor may be variations between representatives of the same species; we thus aim to obtain more specimens, in terms of collecting more samples from the same species but also adding more species. Another limitation stems from the focus of our study. We focus exclusively on circulating sncRNAs in blood cells and thus miss sncRNAs which might be specific to other cell types. In order to reach a comprehensive description of the sncRNAs present in the analysed species, more tissues and specimens will be needed.

CONCLUSION

The detection, annotation and validation of sncRNAs, especially miRNAs, is still a growing field. To understand their function and their potential as biomarkers for diseases, we must first understand how to distinguish actually expressed and valid miRNAs from false positive findings. Conservation is a widely applied feature for identifying miRNAs in related species. Such analyses are often only performed via homology- and sequence-based *in silico* approaches. With our study, we provide a large collection of small RNA NGS expression data for species that have not been analysed before in great detail. We created a comprehensive publicly available online resource for researchers in the field to facilitate the assessment of evolutionarily conserved small RNA sequences.

DATA AVAILABILITY

All sequencing data have been deposited in the Sequence Read Archive with the accession SRP162759.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Edanz (www.edanzediting.com) for editing the English text of a draft of this manuscript.

FUNDING

Saarland University; Saarland; Michael J. Fox Foundation for Parkinson's Research (US) [14446]. Funding for open access charge: Saarland.

Conflict of interest statement. None declared.

REFERENCES

1. Crick,F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.
2. Lee,R.C., Feinbaum,R.L. and Ambros,V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
3. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
4. Veneziano,D., Di Bella,S., Nigita,G., Laganà,A., Ferro,A. and Croce,C.M. (2016) Noncoding RNA: Current deep sequencing data analysis approaches and challenges. *Human Mutat.*, **37**, 1283–1298.
5. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
6. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.-P., Meese,E. and Keller,A. (2018) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
7. Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.

8. Yue,J., Sheng,Y. and Orwig,K.E. (2008) Identification of novel homologous microRNA genes in the rhesus macaque genome. *BMC Genomics*, **9**, 8.
9. Artzi,S., Kiezun,A. and Shomron,N. (2008) miRNAminer: a tool for homologous microRNA gene search. *BMC Bioinformatics*, **9**, 39.
10. Baev,V., Daskalova,E. and Minkov,I. (2009) Computational identification of novel microRNA homologs in the chimpanzee genome. *Comput. Biol. Chem.*, **33**, 62–70.
11. Long,J.-E. and Chen,H.-X. (2009) Identification and characteristics of cattle MicroRNAs by homology searching and small RNA cloning. *Biochem. Genet.*, **47**, 329–343.
12. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stähler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
13. Backes,C., Meese,E. and Keller,A. (2016) Specific miRNA disease biomarkers in blood, serum and plasma: challenges and prospects. *Mol. Diagn. Ther.*, **20**, 509–518.
14. Keller,A., Fehlmann,T., Ludwig,N., Kahraman,M., Laufer,T., Backes,C., Vogelmeier,C., Diener,C., Biertz,F., Herr,C. *et al.* (2018) Genome-wide MicroRNA expression profiles in COPD: Early predictors for cancer development. *Genomics Proteomics Bioinformatics*, **16**, 162–171.
15. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease microRNAs using next-generation sequencing. *Alzheimers Demen.*, **12**, 565–576.
16. Fehlmann,T., Ludwig,N., Backes,C., Meese,E. and Keller,A. (2016) Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol.*, **13**, 1084–1088.
17. Keller,A., Leidinger,P., Bauer,A., Elsharawy,A., Haas,J., Backes,C., Wendschlag,A., Giese,N., Tjaden,C., Ott,K. *et al.* (2011) Toward the blood-borne miRNome of human diseases. *Nat. Methods*, **8**, 841–843.
18. Keller,A., Leidinger,P., Vogel,B., Backes,C., ElSharawy,A., Galata,V., Mueller,S.C., Marquart,S., Schrauder,M.G., Strick,R. *et al.* (2014) miRNAs can be generally associated with human pathologies as exemplified for miR-144. *BMC Med.*, **12**, 224.
19. Kahraman,M., Laufer,T., Backes,C., Schrörs,H., Fehlmann,T., Ludwig,N., Kohlhaas,J., Meese,E., Wehler,T., Bals,R. *et al.* (2017) Technical stability and biological variability in microRNAs from dried blood spots: a lung cancer therapy-monitoring showcase. *Clin. Chem.*, **63**, 1476–1488.
20. Pirritano,M., Fehlmann,T., Laufer,T., Ludwig,N., Gasparoni,G., Li,Y., Meese,E., Keller,A. and Simon,M. (2018) NGS analysis of total small non coding RNAs from low input RNA from dried blood sampling. *Anal. Chem.*, **90**, 11791–11796.
21. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Würstle,M.L., Hübenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.
22. R Core Team (2018) *R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing Vienna*, Austria, https://www.r-project.org.
23. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* , **57**, 289–300.
24. Mangiafico,S. (2019) rcompanion: Functions to Support Extension Education Program Evaluation. *R package version 2.0.10* . https://rcompanion.org.
25. Wickham,H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, NY, https://ggplot2.tidyverse.org.
26. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
27. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
28. Schliep,K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592.
29. Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.
30. Weese,D., Holtgrewe,M. and Reinert,K. (2012) RazerS 3: faster, fully sensitive read mapping. *Bioinformatics*, **28**, 2592–2599.
31. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
32. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
33. Vitsios,D.M., Kentepozidou,E., Quintais,L., Benito-Gutiérrez,E., van Dongen,S., Davis,M.P. and Enright,A.J. (2017) Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Res.*, **45**, e177.
34. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grässer,F. *et al.* (2015) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
35. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grässer,F.A., Lenhof,H.-P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, doi:10.1093/nar/gkz097.
36. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
37. Fromm,B., Domanska,D., Hackenberg,M., Mathelier,A., Hoye,E., Johansen,M., Hovig,E., Flatmark,K. and Peterson,K.J. (2018) MirGeneDB2.0: the curated microRNA Gene Database. bioRxiv doi: https://doi.org/10.1101/258749, 05 February 2018, preprint: not peer reviewed.
38. Meder,B., Backes,C., Haas,J., Leidinger,P., Stähler,C., Großmann,T., Vogel,B., Frese,K., Giannitsis,E., Katus,H.A. *et al.* (2014) Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin. Chem.*, **60**, 1200–1208.
39. Keller,A., Leidinger,P., Vogel,B., Backes,C., ElSharawy,A., Galata,V., Mueller,S.C., Marquart,S., Schrauder,M.G., Strick,R. *et al.* (2014) MiRNAs can be generally associated with human pathologies as exemplified for miR-144. *BMC Med.*, **12**, 224.
40. Muñoz-Culla,M., Irizar,H., Sáenz-Cuesta,M., Castillo-Triviño,T., Osorio-Querejeta,I., Sepúlveda,L., De Munain,A.L., Olascoaga,J. and Otaegui,D. (2016) SncRNA (microRNA & snoRNA) opposite expression pattern found in multiple sclerosis relapse and remission is sex dependent. *Sci. Rep.*, **6**, 20126.
41. Goodarzi,H., Nguyen,H.C., Zhang,S., Dill,B.D., Molina,H. and Tavazoie,S.F. (2016) Modulated expression of specific tRNAs drives gene expression and cancer progression. *Cell*, **165**, 1416–1427.
42. Huang,S.-q., Sun,B., Xiong,Z.-p., Shu,Y., Zhou,H.-h., Zhang,W., Xiong,J. and Li,Q. (2018) The dysregulation of tRNAs and tRNA derivatives in cancer. *J. Experiment. Clin. Cancer Res.*, **37**, 101.
43. Zhou,Y., Goodenbour,J.M., Godley,L.A., Wickrema,A. and Pan,T. (2009) High levels of tRNA abundance and alteration of tRNA charging by bortezomib in multiple myeloma. *Biochem. Biophys. Res. Commun.*, **385**, 160–164.
44. Nicolas,F.E., Hall,A.E., Csorba,T., Turnbull,C. and Dalmay,T. (2012) Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway. *FEBS Lett.*, **586**, 1226–1230.

A mouse tissue atlas of small noncoding RNA

Alina Isakova[a], Tobias Fehlmann[b], Andreas Keller[b,c], and Stephen R. Quake[a,d,e,1]

[a]Department of Bioengineering, Stanford University, Stanford, CA 94305; [b]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany; [c]Department of Neurology, School of Medicine, Stanford University, Stanford, CA 94305; [d]Department of Applied Physics, Stanford University, Stanford, CA 94305; and [e]Chan Zuckerberg Biohub, San Francisco, CA 94158

Contributed by Stephen R. Quake, July 29, 2020 (sent for review February 10, 2020; reviewed by John C. Marioni and Igor Ulitsky)

Small noncoding RNAs (ncRNAs) play a vital role in a broad range of biological processes both in health and disease. A comprehensive quantitative reference of small ncRNA expression would significantly advance our understanding of ncRNA roles in shaping tissue functions. Here, we systematically profiled the levels of five ncRNA classes (microRNA [miRNA], small nucleolar RNA [snoRNA], small nuclear RNA [snRNA], small Cajal body-specific RNA [scaRNA], and transfer RNA [tRNA] fragments) across 11 mouse tissues by deep sequencing. Using 14 biological replicates spanning both sexes, we identified that ∼30% of small ncRNAs are distributed across the body in a tissue-specific manner with some also being sexually dimorphic. We found that some miRNAs are subject to "arm switching" between healthy tissues and that tRNA fragments are retained within tissues in both a gene- and a tissue-specific manner. Out of 11 profiled tissues, we confirmed that brain contains the largest number of unique small ncRNA transcripts, some of which were previously annotated while others are identified in this study. Furthermore, by combining these findings with single-cell chromatin accessibility (scATAC-seq) data, we were able to connect identified brain-specific ncRNAs with their cell types of origin. These results yield the most comprehensive characterization of specific and ubiquitous small RNAs in individual murine tissues to date, and we expect that these data will be a resource for the further identification of ncRNAs involved in tissue function in health and dysfunction in disease.

miRNA | noncoding | sex dimorphism | tissue specificity

Small noncoding RNAs (ncRNAs) are a large family of endogenously expressed transcripts, 18 to 200 nt long, that play a crucial role in regulating cell function (1, 2). Seen mainly as "junk" RNA of unknown function two decades ago, today small ncRNAs are believed to be involved in nearly all developmental and pathological processes in mammals (2–4). While the exact function of many ncRNAs remain unknown, numerous studies have revealed the direct involvement of various small ncRNAs in regulation of gene expression at the levels of posttranscriptional mRNA processing (5–7) and ribosome biogenesis (8). Aberrant expression of small ncRNAs, in turn, has been associated with diseases such as cancer, autoimmune disease, and several neurodegenerative disorders (9, 10).

Mammalian cells express several classes of small ncRNA, including microRNA (miRNA) (11), small interfering RNAs (siRNA), small nucleolar RNAs (snoRNA) (12), small nuclear RNA (snRNA) (13), PIWI-interacting RNA (piRNA) (14), and tRNA-derived small RNAs (tRFs) (15), with some being shown to be expressed in a tissue- (16), cell type- (17), or even cell state-specific manner (18). Through their interactions with ribosomes and mRNA, these small noncoding molecules shape the dynamic molecular spectrum of tissues (4, 17). Despite extensive knowledge of ncRNA biogenesis and function, much remains to be explored about tissue- and sex-specific small ncRNA expression. Given the emerging role of ncRNAs as biomarkers (19, 20) and potent therapeutic targets (21), a comprehensive reference atlas of tissue small ncRNA expression would represent a valuable resource not only for fundamental but also for clinical research.

The first attempts to catalog tissue-specific mammalian small ncRNAs began a decade ago with characterization of miRNA levels (16, 22, 23). While these pioneering microarray-, qPCR-, and Sanger sequencing-based studies mapped only a limited number of highly expressed miRNA, they nevertheless established a "gold standard" reference for the following decade of miRNA research. Efforts to characterize tissue-specific noncoding transcripts have recently resumed with the advent of RNA sequencing (RNA-seq), which greatly advanced the discovery of novel and previously undetected miRNA (24, 25). However, no prior study encompasses a spectrum of mammalian tissues from both female and male individuals, nor includes the other noncoding RNA types that have recently been identified to carry out tissue- and cell type-specific functions (26, 27).

Here, we describe a comprehensive atlas of small ncRNA expression across 11 mammalian tissues. Using multiple biological replicates (n = 14) from individuals of both sexes, we mapped tissue-specific as well as broadly transcribed small ncRNA attributed to five different classes and spanning a large spectrum of expression levels. Our data reveal that tissue specificity extends to ncRNA types other than miRNA and provide insights on the tissue-dependent distribution of miRNA arms and tRNA fragments. We have also discovered that certain miRNAs are broadly sexually dimorphic, while other show sex bias in the context of specific tissues. Finally, integrating our ncRNA expression measurements with the scATAC-seq data (28) enabled us to map cell type specificity of small ncRNA expressed in the adult mouse brain.

Results

Small ncRNA Expression Atlas of Mouse Tissues. We profiled the expression of small ncRNA across 10 tissues from adult female (n = 10) and 11 from adult male (n = 4) C57BL/6J mice (Fig. 1*A* and Dataset S1). We generated a dataset comprising a total of 140 small ncRNA sequencing libraries from brain, lung, heart, muscle, kidney, pancreas, liver, small intestine, spleen, bone

Significance

We report a systematic unbiased analysis of small RNA molecule expression in 11 different tissues of the model organism mouse. We discovered uncharacterized noncoding RNA molecules and identified that ∼30% of total noncoding small RNA transcriptome are distributed across the body in a tissue-specific manner with some also being sexually dimorphic. Distinct distribution patterns of small RNA across the body suggest the existence of tissue-specific mechanisms involved in noncoding RNA processing.

Author contributions: A.I. and S.R.Q. designed research; A.I. performed research; A.I., T.F., and A.K. analyzed data; and A.I. and S.R.Q. wrote the paper.

Reviewers: J.C.M., European Molecular Biology Laboratory–European Bioinformatics Institute; and I.U., Weizmann Institute of Science.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: steve@quake-lab.org.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2002277117/-/DCSupplemental.

First published September 25, 2020.

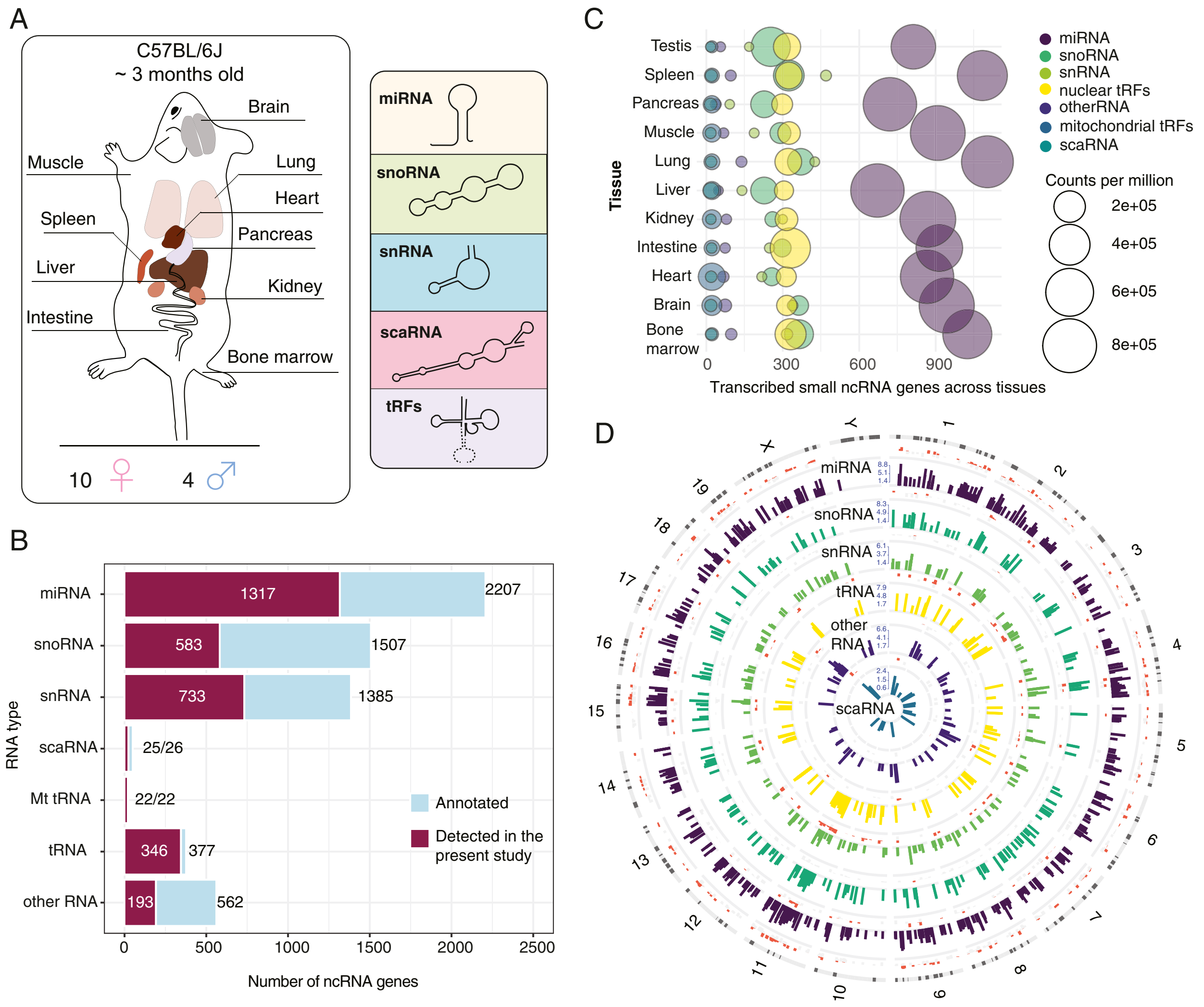


Fig. 1. Small ncRNA expression across mouse tissues. (*A*) Tissues and ncRNA classes profiled in the current study. Ten somatic tissues were collected from adult mice (*n* = 14). Testes were collected from male mice (*n* = 4). (*B*) ncRNAs identified in the current study. Numbers indicate detected and total annotated within GENCODE M20 miRNA, snoRNA, snRNA, scaRNA, Mt_tRNA, as well as high-confidence tRNA listed in GtRNAdb. (*C*) Coverage of ncRNA types within the profiled tissues. ncRNA was considered transcribed in a tissue if detected at >1 cpm. (*D*) Genomic map of small RNAs (sRNAs) expression across mouse genome. The bars show the log-transformed normalized expression count of ncRNAs. The red and gray bars around each circle represent the variance of each sRNA across 10 mouse tissues. Red denotes highly (the SD of expression above 25% of the mean value), and gray, low, variable ncRNAs (SD below 25% of the mean value).

marrow, and testes RNA. Each library yielded ∼5 to 20 million reads mapping to the mouse genome, out of which, on average ∼7 million mapped to the exons of small ncRNA genes (*Materials and Methods*), resulting in a total of ∼100 million ncRNA reads per tissue (*SI Appendix*, Fig. S1*A*). Using the GENCODE M20 (29), GtRNAdb (30), and miRBase (31) mouse annotations, we mapped the expression of distinct small ncRNA classes: miRNA, snRNA, snoRNA, scaRNA, tRF, and other small ncRNA in profiled tissues (Fig. 1*B*). Among all of the tissues, we identified a total of 1,317 distinct miRNA, 733 snRNA, 583 snoRNA, 25 scaRNA, 346 tRNA, 22 mitochondrial tRNA, and 193 other miscellaneous small ncRNA genes, which correspond to 60%, 53%, 39%, 96%, 92%, 100%, and 34% of annotated transcripts of each respective class (Fig. 1*B*). miRNA was the most abundant small ncRNA type in our libraries, followed by snoRNA, snRNA, and tRFs (Fig. 1*C* and *SI Appendix*, Fig. S1*B*). snoRNA and snRNA are believed to yield incomplete recovery in small RNA-seq experiments due to their secondary structure (32). With respect to protein coding genes, the detected tRFs were intergenic, snoRNAs were of intronic origin, snRNAs and scaRNAs were intronic and intergenic (63/35% and 64/36%, respectively), and miRNAs were transcribed from either introns (53%), exons (11%), or intergenic regions (11%) (*SI Appendix*, Fig. S1*C*). The number of distinct ncRNA greatly varied across tissues; for example, lymphoid tissues (lung, spleen, and bone marrow) contained the largest number of distinct ncRNA, while pancreas and liver contained the fewest (Fig. 1*C* and *SI Appendix*, Fig. S1*B*). Furthermore, within the profiled tissues, we detected 95.1% of miRNA precursors denoted by miRBase v22 database as high-confidence transcripts (31). Using these data, we have reconstructed genome-wide expression map of various small ncRNA types across 11 murine tissues (Fig. 1*D* and Dataset S2).

Tissue-Specific Expression of Small ncRNA. We first assessed the differences in the levels of small ncRNAs across profiled tissues at the gene level, based on the expression of all assayed RNA types. Dimensionality reduction via *t*-distributed stochastic neighbor embedding (*t*-SNE) (*Materials and Methods*) on ncRNA genes revealed a robust clustering of samples according to tissue types (Fig. 2*A*). For each ncRNA, we have computed the tissue specificity index (TSI), as described previously in ref. 33. We observed that ~17% of all detected ncRNA were present in only one tissue (TSI = 1) (*SI Appendix*, Fig. S2*A*), while the remaining ncRNA were either ubiquitously expressed or found in some but absent in other tissues. We next ran a differential gene expression (DGE) analysis on all detected ncRNAs across 11 tissues (*Materials and Methods*) and found that out of 3,219 detected genes, 897 (28%) contribute to the tissue-specific signature of ncRNA expression (at a false discovery rate [FDR] < 1%) (Fig. 2 *B* and *C* and Dataset S3). Interestingly, we found brain to contain the highest number of unique transcripts not present in other tissues (~400) (Fig. 2 *B* and *C* and Dataset S3) even though lung, spleen, and bone marrow expressed the widest spectrum of detected genes (Fig. 1*C*). We found miRNA to be the main contributor of tissue specificity (reflected by the fraction of each specific RNA type with TSI > 0.9) (Fig. 2 *B* and *C*); however, we have also identified hundreds of ncRNAs of other types that are expressed in a tissue-specific fashion (*SI Appendix*, Fig. S2).

Tissue-Specific snoRNAs. We found that snoRNA alone is capable of separating the majority of profiled tissues based on their transcript levels (*SI Appendix*, Fig. S2*B*), with over 200 snoRNA showing tissues-specific patterns (*SI Appendix*, Fig. S2*C* and Dataset S3). For example, we discovered that maternally imprinted *AF357428* (also known as MBII-78), *AF357341* (MBII-19), and *Gm25854*, transcribed from a 10-kb region of chromosome 12qF1, are up-regulated in the brain and muscle. Interestingly, two other snoRNA, *Gm22962* and *Gm24598*, followed the same tissue-specificity pattern, despite being transcribed from other chromosomes (9qC and XqA7.1, respectively). While present at low levels, we also identified *Snora35* (MBI-36) and Snord116 (MBII-85), known to be involved in neurodevelopmental disorders (34), to be brain exclusive (TSI = 1). We observed high levels of *Snord17* and *Snord15a* in the spleen and bone marrow, and lower levels in other tissues. These snoRNAs have been previously reported among up-regulated genes in bacterial infection of soft tissues (35), suggesting the association of these transcripts with immune cells. We found several snoRNAs present mainly in the pancreas, such as *Snord123* (*SI Appendix*, Fig. S2*C*), located 3 kb upstream of the pancreatic cancer-associated *Sema5a* gene, and *Gm22888* (Fig. 2*B*) located within the introns of *Ubap2*. We also identified a large number of other snoRNA, the exact function of which is still unknown, to be enriched in either one or multiple tissues (*SI Appendix*, Fig. S2*C*), among which are *Snord53*, *Gm24339*, *Gm26448*, *Snora73a*, *Snord104* in lymphoid tissues, and *Snord34*, *Gm24837* in testes. Finally, we show that some snoRNAs, such as *Snord70* and *Snord66*, which are often used as normalization controls in qPCR-based assays (36, 37), are also expressed in a tissue-biased manner (*SI Appendix*, Fig. S2*C*).

Tissue-Specific Expression of *Rny*, *Terc*, and Other ncRNA. Analyzing the levels of other ncRNA classes, we found that the brain contains high levels of *Rny1* compared to other tissues (Fig. 2*B*). We also observed that the levels of another transcript from the same class, *Rny3*, are elevated in pancreas, brain, and kidney (*SI Appendix*, Fig. S2*D*). The precise biological function of *Rny1* and *Rny3* is so far undefined, although they have been suggested to maintain RNA stability (38).

Interestingly, we detected the presence of telomerase RNA component (*Terc*) in analyzed somatic tissues, with the highest levels seen in the bone marrow and spleen. Together with previous reports that identify telomerase activity in hematopoietic cells (39) and show *Terc*+ cells to secrete inflammatory cytokines (39, 40), our data suggest that *Terc* is specific to cells of hematopoietic origin. Among other ncRNA types that we found to be differentially expressed across profiled tissues are snRNA and scaRNA, both known to be involved in the regulation of splicing events (41). We observed that, similarly to primate orthologs (42), mouse Rnu11 and Scarna6 are preferentially found in lymphoid tissues, while three snRNAs of unverified function, *Gm25793*, *Gm22448*, and *Gm23814*, are specific to the brain (*SI Appendix*, Fig. S2*D*).

Tissue-Specific miRNA. We found ~400 miRNAs differentially expressed across profiled tissues (Dataset S3). Out of these 400, nearly one-quarter are specific to the brain, with some being uniquely expressed within the tissue (*SI Appendix*, Fig. S3*A*). We identified both well-described brain-specific miRNAs, such as *mir-9*, *mir-124*, *mir-219*, *mir-338* (16, 24, 33), as well as those which are missing from existing catalogs, such as *mir-666*, *mir-878*, *mir-433*, etc. (Fig. 2*B* and *SI Appendix*, Fig. S3*A*). Examples of other miRNAs previously unknown to be tissue-specific include *mir-499* in the heart, *mir-3073b* in the kidney, and *mir-215* and *mir-194* in the intestine (*SI Appendix*, Fig. S3*A*). We also observed multiple miRNAs present in several tissues but absent in others, reflecting the cellular composition of the tissues. Surprisingly, we also found a few miRNAs, such as *mir-134*, *mir-182*, *mir-376c*, *mir299a*, *mir-3061*, and *mir-7068* (*SI Appendix*, Fig. S3*A*) to be shared solely between muscle, brain, and pancreas, which, in turn, do not contain any evident common cell types that are absent in other tissues. Independently, unsupervised clustering of the top 400 most differentially expressed ncRNA in our dataset also revealed that two out of three identified clusters comprise ncRNA genes that are up-regulated in the above-mentioned three tissues (*SI Appendix*, Fig. S3*B*). Altogether, these findings suggest that certain small ncRNAs are involved in maintaining a specific function within the brain, pancreas, and muscle, which could, for example, be ion transport or exocytosis.

Tissue-Specific Arm Selection of miRNA. Assessing the overall abundance of 5p or 3p arms of miRNA across tissues, we found no significant bias in strand selection (*SI Appendix*, Fig. S4*A*). For many miRNAs, we generally observed the dominance of either 3p or 5p arm, while for some we also detected high levels of both arms present in one or multiple tissues (Fig. 2 *C* and *D* and Dataset S4). Nonetheless, we found that ~5% of all miRNAs switch their arm preference between tissues. Some of them, like *mir-337*, *mir-106b*, and *mir-26b*, are represented by both arms in certain tissues but only by one of the arms in other (Fig. 2*D* and Dataset S4). More striking examples of complete arm switching from one tissue to another are *mir-141* and *mir-350* (Fig. 2*D*). *miR-141-5p* but not *-3p* is present in the brain, and *-3p* but not *-5p* in the testes, while both arms are found in the pancreas and intestine. In the case of *mir-350*, both arms are detected in the bone marrow, brain, and spleen, while only the 5p arm is present in the heart, lung, and muscle, and the 3p arm in the pancreas (Fig. 2*D*). This highlights the complexity of tissue-dependent miRNA biogenesis and indicates that the phenomena of miRNA arm switching, so far only observed in cancer, extends to healthy mammalian tissues (43–45).

Ubiquitous ncRNA Transcripts. We detected many ubiquitously expressed ncRNAs across tissues (*SI Appendix*, Fig. S4*B*). Among these are ncRNAs known to be expressed in a large number of cell types, such as *let-7d-3p*, *miR-320-3p* (25), ncRNAs

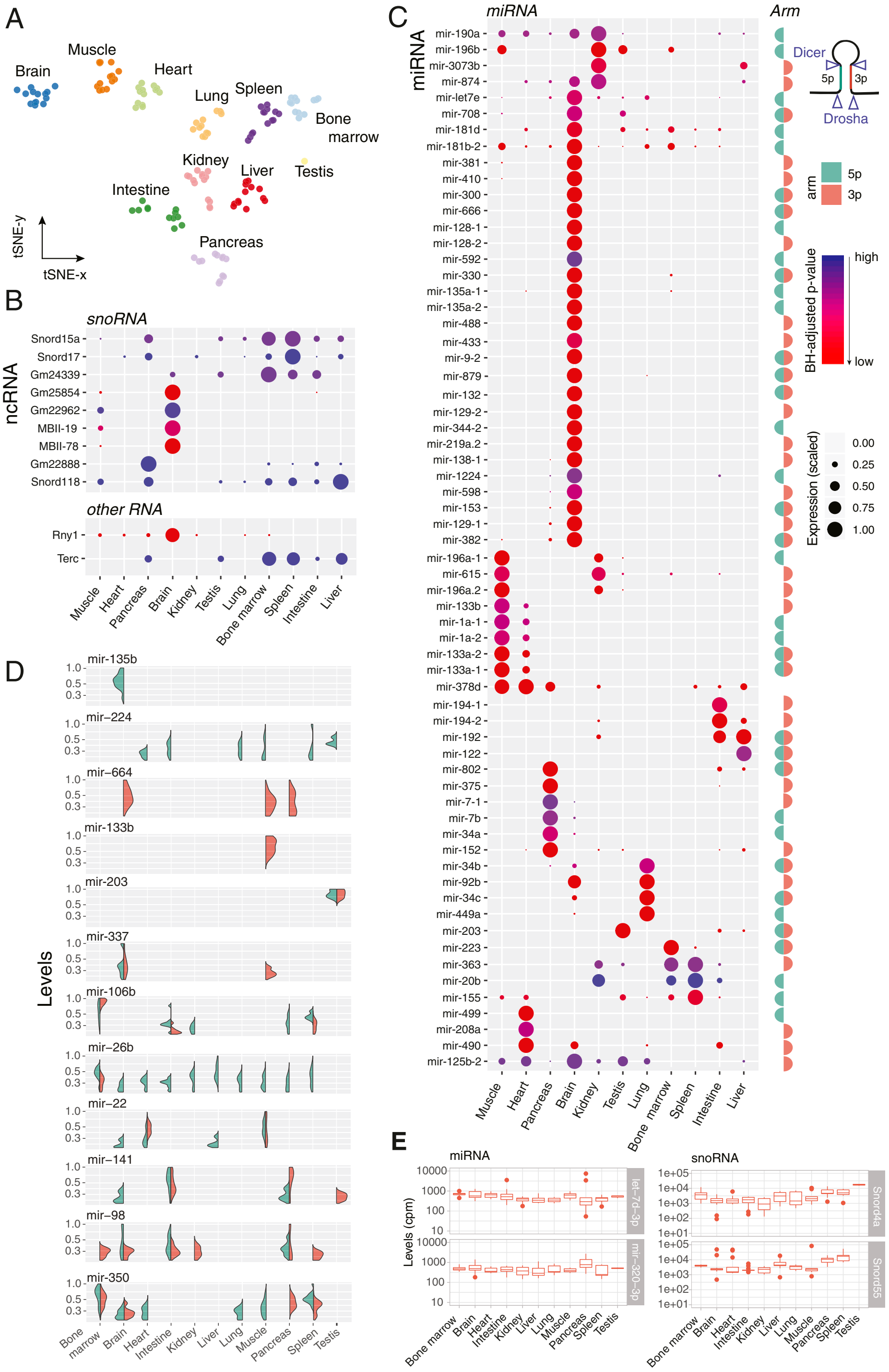


Fig. 2. Tissue-specific patterns of small ncRNA expression. (*A*) *t*-SNE projection of ncRNA expression patterns performed on ∼4,000 ncRNA genes of various classes detected in 11 mouse tissues. (*B*) Dot plot of tissue-specific snoRNA, *Rny1* and *Terc*, identified as the most tissue-specific (Benjamini–Hochberg [BH]-adjusted *P* value < 0.01 in LRT test). The size of the dot represents scaled log-transformed normalized counts. (*C*) Dot plot of tissue-specific miRNA. Only a subset of miRNA passing the specificity threshold (BH-adjusted *P* value < 0.01 in LRT test) is shown. "Arm" column denotes whether *5p-*, *3p-*, or both arms are passing the specificity threshold. (*D*) Levels of miRNA arms detected across tissues. Representative miRNAs, for which we consistently detect either one of the arms, both, or switched arm between tissues. The *y* axis represents normalized scaled counts. (*E*) Examples of ubiquitous ncRNA present in all tissues.

the cell type specificity of which is still unknown, such as *Snord4a* and *Snord55*, as well as those known to be expressed in the cell types that are abundant in all tissues (like endothelial *miR-151-5p*) (*SI Appendix*, Fig. S4*B*).

Novel miRNAs. We have recently demonstrated that the current miRbase annotation of mammalian miRNA remains incomplete but can be readily expanded with the help of emerging small RNA-seq data (46). To search for novel miRNA in our data, we first processed all unmapped reads using miRDeep2 (47) and selected 473 genomic regions harboring a putative miRNA gene supported by at least 10 sequencing reads. To refine this list, we employed three parallel strategies: 1) we searched for the presence of the putative miRNA in 141 public Argonaute CLIP-seq (AGO-CLIP) datasets from various mouse cell and tissue types; 2) we performed a literature and database search for prior mentions of the putative miRNA; and 3) we ran an RT-qPCR validation of selected candidates. Analysis of AGO-CLIP data showed evidence for 214 out of 473 candidates (total, >5 counts). We also found that 87 out of 473 novel miRNA were previously reported within other studies (48–52) (Fig. 3*A*). Importantly, 52 novel miRNAs identified by this and previous studies were not present in the AGO-CLIP data (Dataset S5). The RT-qPCR quantification of two miRNAs selected from this list, *17_11530* and *7_16137*, in turn, confirmed the existence of these transcripts (*SI Appendix*, Fig. S5*A*). On the other hand, we identified novel miRNAs (*17_8620*, *4_6440*, *9_15723*) that are supported by AGO-CLIP data, prior reports, or both, but for which we could not confirm the existence through RT-qPCR (*SI Appendix*, Fig. S5*A*). We also found a novel miRNA that, among the three validation methods, was only verified through RT-qPCR. Interestingly, the genomic coordinates of this miRNA, *14_6588*, matched the coordinates of another, annotated one, *mir-802a*. Unlike *mir-802a*, however, *14_6588* is transcribed from the negative DNA strand and is only present in the brain (*SI Appendix*, Fig. S5*B*). Altogether, by comparing the miRNA levels measured through RT-qPCR with the tissue transcript abundance identified by small RNA-seq, we validated 12 novel miRNAs that were either also reported by others (Fig. 3*B* and *SI Appendix*, Fig. S5*A*) or uniquely identified in the present study (Fig. 3 *C* and *D* and *SI Appendix*, Fig. S5*A*).

We found that the majority of putative miRNAs are present in only one tissue (312), but a small number (4) are found in all 11 tissues (Fig. 3*E*). Principal-component analysis on the newly identified miRNAs, supported by at least 50 reads, showed a clear separation of brain, lung, and muscle from other tissues based on expression values. Similar to annotated transcripts, novel miRNAs demonstrate a spectrum of tissue specificity with some being ubiquitously expressed, while others are only present in one tissue (*SI Appendix*, Fig. S5*C*). Differential expression analysis on putative novel miRNAs identified six miRNAs to be also expressed in a sex-specific manner. Strikingly, all six were male-dominant, with one of them even found to be consistently up-regulated in two tissues, male muscle and pancreas (*SI Appendix*, Fig. S5*D*).

Tissue-Resident tRNA Fragments. About a quarter of our small RNA-seq libraries consisted of tRFs—fragments of either mature

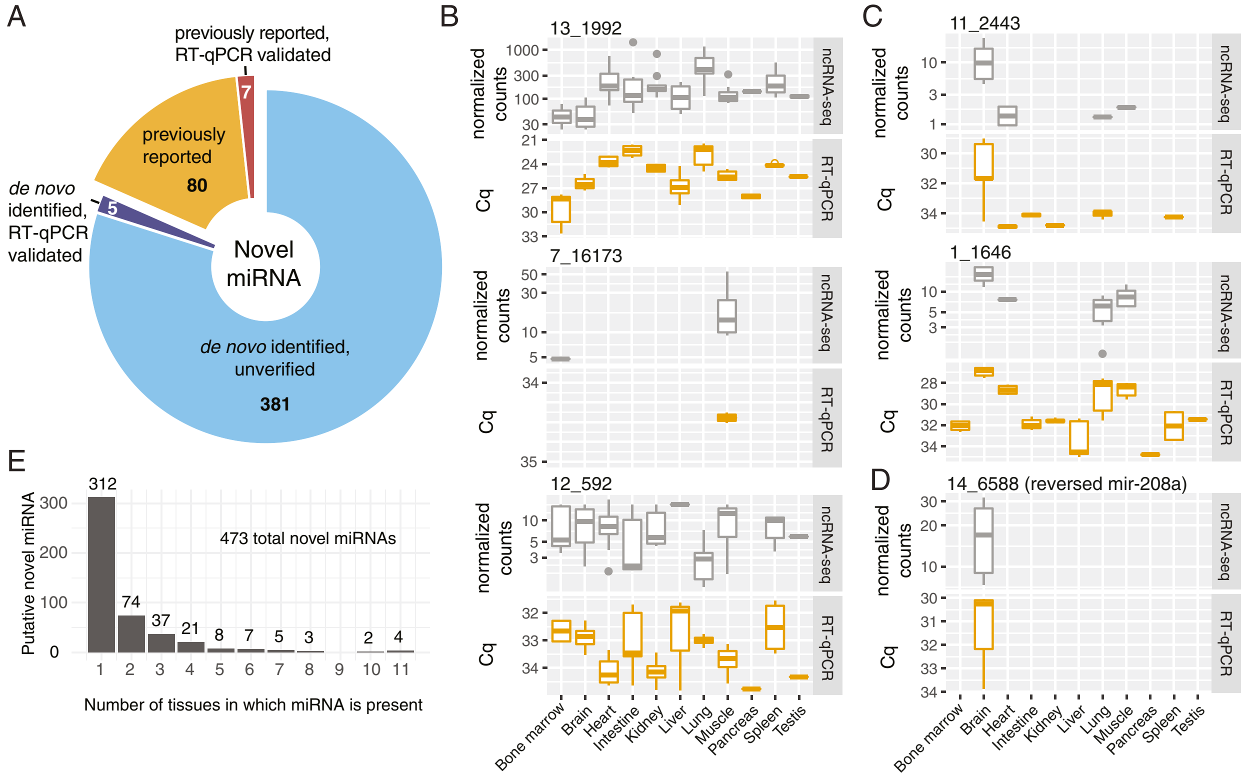


Fig. 3. miRNAs uniquely detected in the present study. (*A*) Pie chart of predicted and verified miRNA. (*B*) Examples of RT-qPCR verified miRNAs identified by this and previous studies. The levels of miRNA were determined from small RNA-seq data (DESeq2 normalized counts) and through RT-qPCR (Cq, adjusted for the sample-to-sample variability using cel-mir-39 spike in control). (*C*) Same as *C* but for miRNAs detected uniquely within the present study. (*D*) RT-qPCR verified miRNA, 14_6588, transcribed from the negative strand of mir-208a. (*E*) Tissue specificity of putative miRNA.

of precursor tRNA molecules enzymatically cleaved by angiogenin (Ang), Dicer, RNaseZ, and RNaseP (7, 53) (Fig. 4*A*).

"Exact tRNA multimapping" of these fragments to the mouse genome revealed the presence of tRFs of various sizes. Interestingly, consistent with a previous report on tRFs in human cell lines (7), we observed a major difference in the size of fragments originating from either nuclearly or mitochondrially encoded tRNA (ntRNA and mtRNA, respectively). While the majority of ntRNA fragments were 33 nt long, mtRNA fragments spanned a large size range of 18 to 54 nt (Fig. 4*B* and *SI Appendix*, Fig. S6*A*). This distinct pattern of fragment sizes reflected the bias in the amounts of tRF types originating from nt- and mtRNA (Fig. 4*C* and *SI Appendix*, Fig. S6*B*). We observed that the distribution of nuclear tRFs was largely skewed toward 5′tR-halves, generated by the cleavage in the anticodon loops of mature tRNA. However, within mitochondrial tRFs, we identified a more uniform representation of cleaved fragments. Furthermore, we found that the relative abundance of tRF types, within both ntRNA and mtRNA space, is not constant, but varies across tissues (Fig. 4 *C* and *D*). The tissue-type differences are also present across different tRNA isoacceptors and even its anticodons (Fig. 4*D* and *SI Appendix*, Fig. S6*B*). In the case of nuclear tRFs, the vast majority of fragments in each tissue were attributed to glycine, glutamine, valine, and lysine tRNA, with the intestine containing the largest amounts of the respective 5′tR-halves. Since the abundance of these specific fragments has been shown to correlate with the levels of functional angiogenin in the cell (54), we speculate that the biological explanation of the intestine yielding high levels of tRFs is due to the activity of Ang4, one of the five Ang proteins in mouse, highly expressed in Paneth and Goblet cells of the intestinal epithelium (55, 56).

For many ntRFs, the distribution between 3′- and 5′-, tRF and tR-halves was surprisingly shifted toward one form, i.e., one fragment type was present at higher amounts than others (Fig. 4 *E* and *F*). For the majority of fragments, we found 5′tR- or 3′tR-halves to be the most dominant fragment type. However, in the rare cases, we found 5′- or 3′tRFs to dominate other fragments in a tissue-specific manner. An example of such a fragment is 5′tRF Glu-TTC, which we found to be enriched in the pancreas, compared to other tissues that mostly contained Glu-TTC 5′tR-halves. mtRFs followed a similar trend of fragment shift. We found 5′tRFs of proline-transferring mt-Tp in the heart and 5′tRFs of asparagine-transferring mt-Tn in the liver, while within other tissues we detected different fragment types of these tRNAs (Fig. 4*F*).

miRNAs Are Expressed in a Sex-Specific Manner. Several groups have observed a sex bias in the levels of miRNA in blood, cancer tissues, and human lymphoblastoid cell lines (57, 58). To investigate whether this phenomenon extends to healthy tissues, we compared the ncRNA levels within each tissue coming from either female or male mice. Among ∼6,000 genes assigned to various ncRNA classes, we identified several miRNAs to be differentially expressed between females and males (at FDR < 0.01) (Fig. 5*A*). Some of them are globally sexually dimorphic, while the majority are sex-biased only within a specific tissue. In each somatic tissue, except pancreas, we identified at least two miRNAs differentially expressed between sexes (log2FoldChange > 1, normalized counts > 100, FDR < 0.01) (*SI Appendix*, Fig. S7 *A* and *B*). Kidney and lung contained the highest number of sex-biased miRNAs (27 and 18, respectively), while only two were detected in the heart, five in the muscle, and seven in the brain (Fig. 5*B* and *SI Appendix*, Fig. S7*A*). Three out of eight female-dominant miRNAs: *mir-182*, *mir-148a*, and *mir-145a*, were also shown previously to be estrogen regulated (59), while another miRNA, *mir-340*, was reported to be down-regulated in response to elevated androgen levels (60). Interestingly, we also found that four out of five male-specific miRNAs in the brain are transcribed from a 5-kb region of imprinted Dlk1-Dio3 locus on chromosome 12 (Fig. 5*C*).

Given the innate ability of miRNA to lower the levels of target mRNA (61), we hypothesized that the levels of protein-coding transcripts targeted by sex-biased miRNA would also differ across male and female tissues. To test this hypothesis, we correlated the expression of sex-biased miRNAs with the levels of their respective target mRNAs across profiled tissues (*Materials and Methods*). Among the 60 anticorrelated targets ($r_s < -0.8$, FDR < 0.1) we identified two genes previously shown to be sexually dimorphic (*SI Appendix*, Fig. S7*C*). Specifically, we found *miR-423*, up-regulated in male lung and bone marrow, to negatively correlate with its target, estrogen-related receptor gamma (*Esrrg*) ($r_s = -0.9$, FDR < 0.1), and female-specific *miR-340* to negatively correlate with androgen-associated ectodysplasin A2 receptor (*Eda2r*).

ncRNA-Based Tissue Classification. It is natural to wonder whether the observed variation in ncRNA expression across tissues (Fig. 2 *B* and *C*) would be sufficient to accurately predict the tissue type based solely on small RNA-seq data. To address this question, we set out to construct an algorithm that can learn the characteristics of a healthy tissue from the data reported in the current study and make predictions on new datasets. We limited our analysis to miRNA, since high-throughput tissue data for other ncRNA types is not available. We trained a support vector machine (SVM) model (62) on datasets generated in this study, each containing the expression scores for 1,973 miRNAs (*SI Appendix*, Fig. S8*A*). As a validation dataset, we used available miRNA-seq data released by the ENCODE consortium for multiple mouse tissues (63). Notably, the ENCODE datasets contained data generated for the postnatal and embryonic life stages, as opposed to the adult stage profiled in the current study (Dataset S6). Nonetheless, our SVM model accurately classified postnatal forebrain, midbrain, hindbrain, and neural tube as brain tissue, as well as accurately inferred the tissue types for heart, intestine, kidney, liver, and muscle samples, yielding an overall accuracy of 0.96 (*Materials and Methods*) (see Fig. 6*A* for a full list of accurately predicted tissues as well as for false positives/negatives). For the embryonic tissues, however, our model was able to only reach an accuracy of 0.69. This was mainly due to inability of the model to correctly classify liver tissues instead of assigning them to bone marrow (Fig. 6*A*). Strikingly, in this case, our model accurately predicted the hematopoietic composition of the organ, known to shift from the liver at the embryonic stages to the bone marrow in adulthood (64), rather than the tissue type itself. Furthermore, we identified hematopoiesis-associated *miR-150* and *miR-155* (65) to have highest weights among the features defining the bone marrow in our model (*SI Appendix*, Fig. S8*B*).

We next asked how the identified tissue expression patterns compare to those of individual cell types. To investigate this, we correlated our data with the miRNA data generated for primary mammalian cells by FANTOM5 consortium (25). Comparing mouse samples first, we found that FANTOM5 embryonic and neonatal cerebellum tissues strongly correlated with our brain samples ($r_s = 0.89$ to 0.9), while erythroid cells had the strongest correlation with spleen and bone marrow ($r_s = 0.93$) (*SI Appendix*, Fig. S8*C*). To perform a comparison with human samples, we focused on the expression scores of 531 orthologs detected in both the current study and the FANTOM5 samples (*SI Appendix*, Fig. S8*D*). Spearman correlation coefficients reflected the cell-type composition of tissues (*SI Appendix*, Fig. S8*E*). As such, we observed that mouse bone marrow and spleen had the highest correlation with human B cells, T cells, dendritic cells, and macrophages ($0.5 < r_s < 0.6$), muscle correlated the most with myoblasts and myotubes ($r_s = 0.47$), while brain correlated best with neural stem cells, spinal cord, and pineal and pituitary glands ($r_s = 0.49$) (*SI Appendix*, Fig. S8*E*).

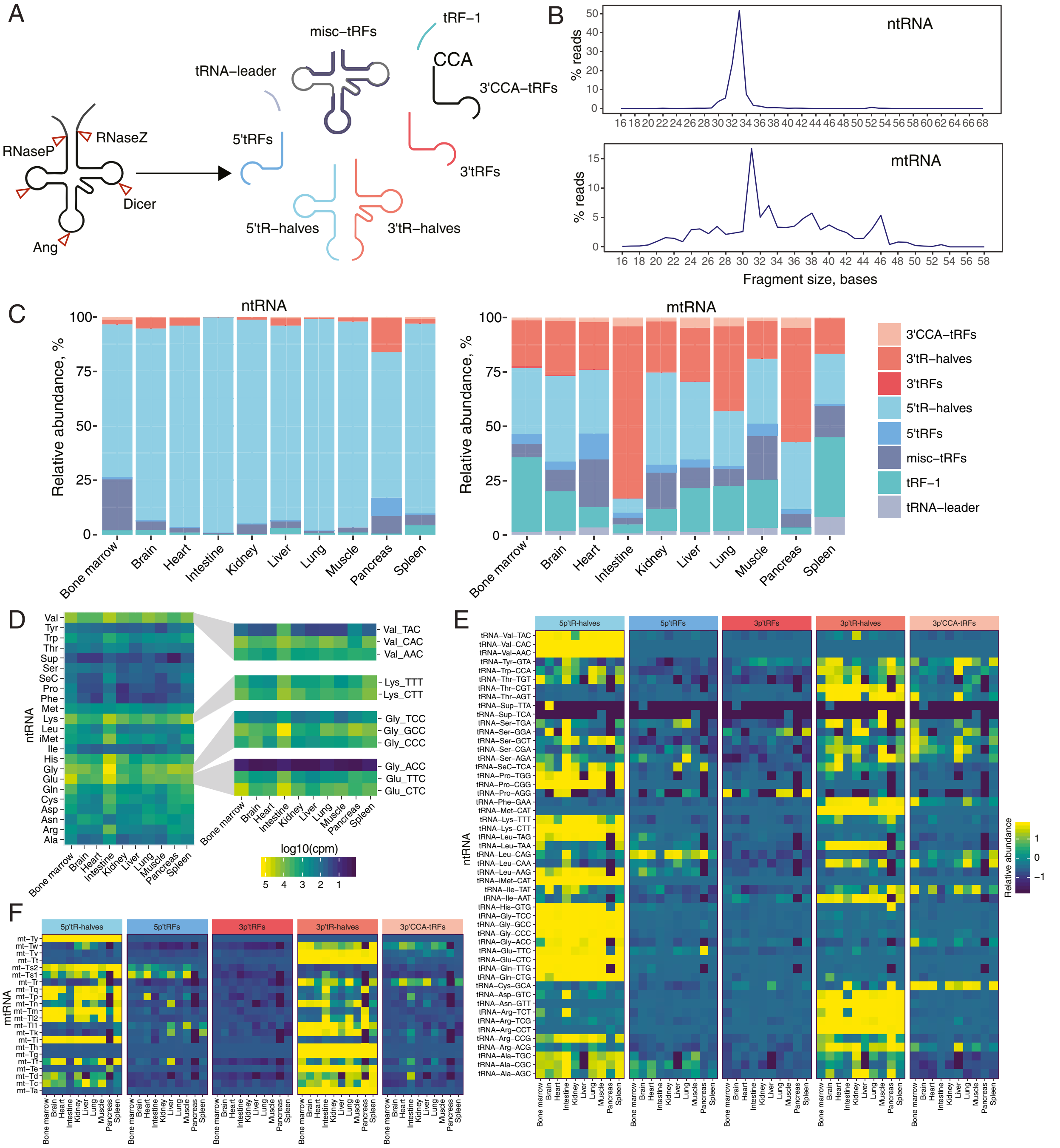


Fig. 4. tRFs detected in mouse tissues. (*A*) Schematic depiction of tRNA cleavage and the resulting fragments. (*B*) Average tRF length identified across tissues for either ntRNA-derived (*Top*) or mtRNA-derived (*Bottom*) fragments. (*C*) Fragment type abundance across tissues. (*D*) Heatmap of tRF levels in tissues. For each of the 23 tRNA types, the sum among its tRFs is plotted. (*E*) Heatmap of relative abundance of ntRNA fragment types (5′-tR-halves, 3′-tR-halves, 5′-tRFs, 3′-tRFs, or 3′-CCA-tRFs) for each tRNA isoacceptor across 10 somatic tissues. Relative abundance is represented by row-wise scaled fractionated scores of tRFs computed by *unitas*. (*F*) Same as *E* but for mtRNA.

Integration of Small RNA-Seq and scATAC-Seq Data. Finally, to deconvolute the complex noncoding tissue profiles and identify the cell types that contribute the observed tissue-specific ncRNAs, we integrated the sequencing data generated within our study with a previously published single-cell ATAC-seq atlas—a catalog of single-cell chromatin accessibility profiles across various cell types (28). First, we compared the expression scores predicted through ATAC-seq measurements with our estimates of ncRNA expression, derived from small RNA-seq. Using the top 400 ncRNAs identified in our analysis as tissue-specific, we correlated average ATAC-seq activity scores and ncRNA levels across eight tissues for which both types of data were available. We

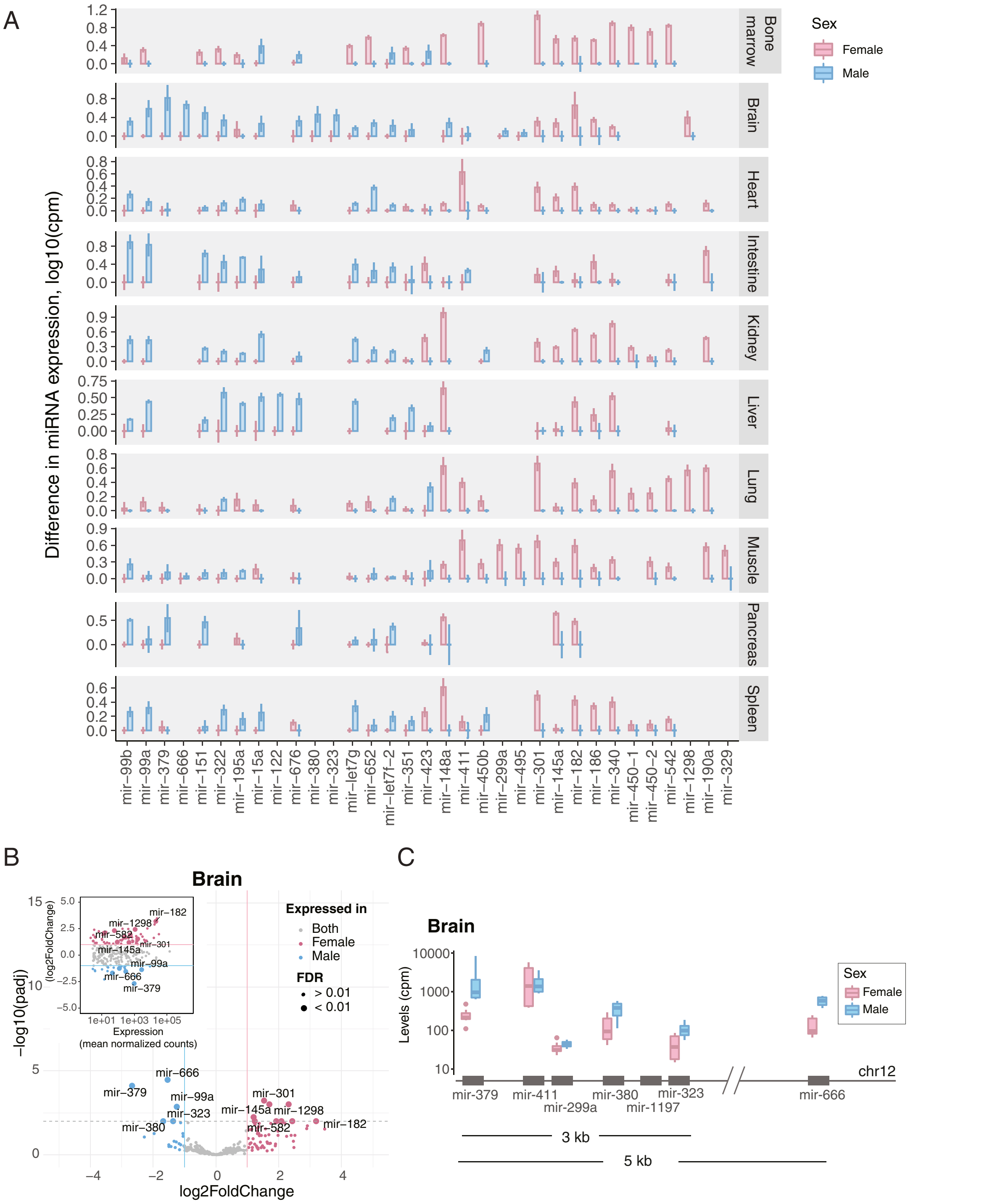


Fig. 5. Sex-specific miRNA expression. (*A*) Sex-dimorphic miRNAs identified in this study. The *y* axis represents the $\log_{10}$ of the difference between mean miRNA levels computed for males and females in each tissue (in counts per million [cpm]). Error bars denote SD of miRNA levels across a tissue within each sex. (*B*) Volcano plot showing miRNAs differentially expressed between the female and the male brain. (*C*) Expression and genomic location of male-biased miRNA in the brain.

observed a strong correlation of both measurements for the brain, liver, and heart, and a weaker correlation for kidney and mixed scores between spleen, bone marrow, and lung (Fig. 6*B*). We found, however, that within each tissue we could attribute the cell type of origin to a number of identified tissue-specific ncRNA. For example, in agreement with previous studies, we could identify that muscle-specific *mir-133a-2* is expressed in cardiomyocytes, while *mir-148a* is expressed in hepatocytes and duct

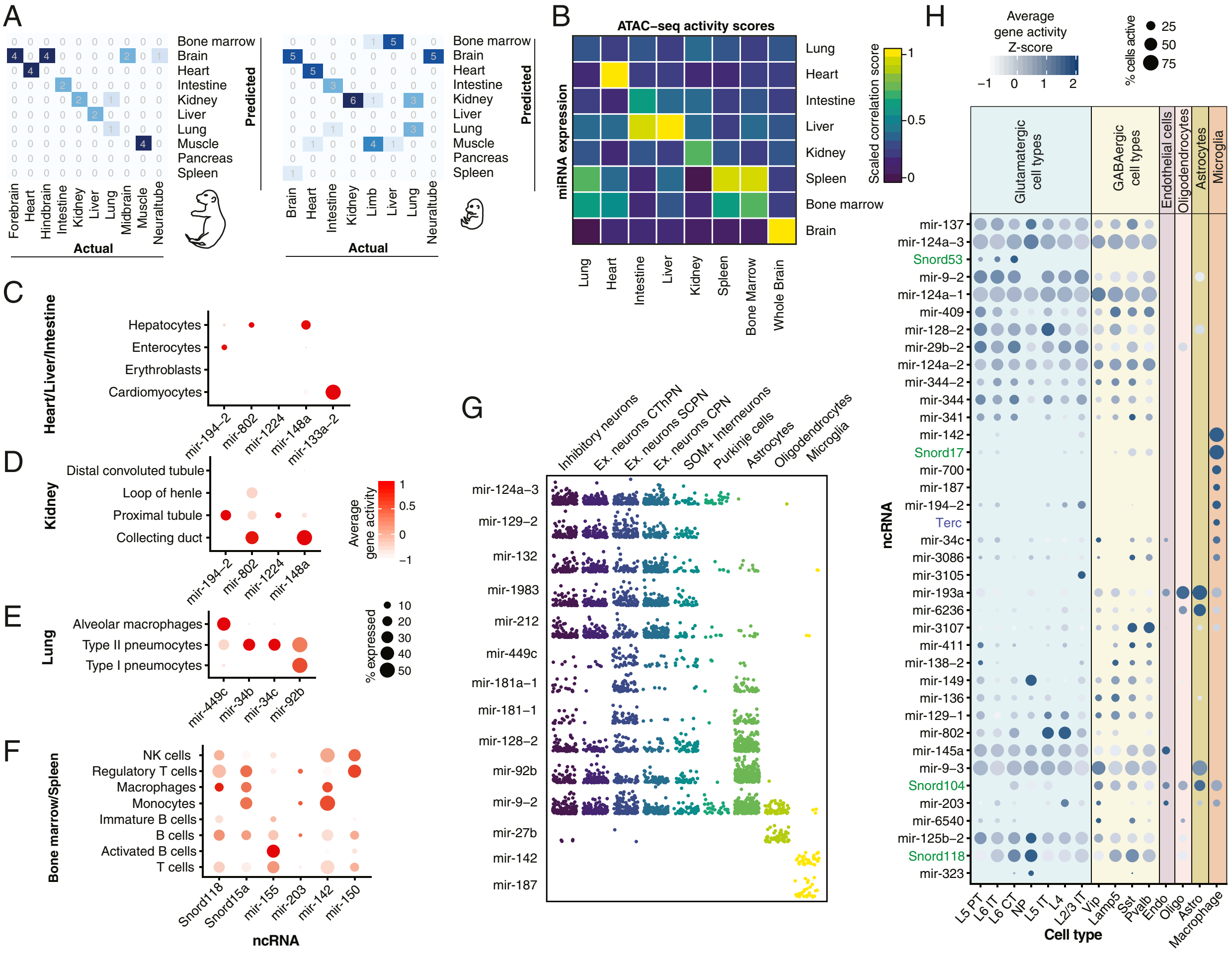


Fig. 6. (*A*) Confusion matrices obtained from SVM tissue classifier for postnatal and embryonic datasets. (*B*) Correlation between small RNA-seq scores derived for top 400 tissue-specific ncRNA genes with the respective activity scores obtained from mouse ATAC atlas. (*C–F*) Average gene activity scores of tissue-specific ncRNAs within each cell type resident to the respective tissue. Gene activity scores for the ncRNAs of interest were retrieved from the mouse ATAC atlas. (*G*) Log$_{10}$ gene activity scores of brain-specific miRNA across individual cells. (*H*) Gene activity scores computed for brain-specific ncRNA from the droplet-based scATAC-seq data (10XGenomics; *Materials and Methods*).

cells (Fig. 6 *C* and *D*) and *mir-194–2* comes from the enterocytes in the gut (24, 25). In addition, we found that in the lung, *mir-449c* in expressed in alveolar macrophages, *mir-34b* and *mir-34c* in type II pneumocytes (Fig. 6*E*), and *mir-155* is found in B cells and even correlates with its maturation status (Fig. 6*F*). Among brain-specific ncRNA, we identified *mir-187* and *mir-142* to be expressed in microglia, *mir-27b* in oligodendrocytes, and *mir-124a-3*, *mir-1983*, *mir-212*, and others in neurons (Fig. 6*G*). For the majority of brain-specific ncRNA identified in our study, however, due to resolution limitations of the mouse ATAC atlas data, we were unable to unambiguously define the cell type of origin. To overcome that, we analyzed a complimentary single-cell ATAC-seq dataset, generated specifically from the mouse adult brain (obtained from 10X Genomics; *Materials and Methods*) and mapped the activity of the brain-specific ncRNA to 15 cell types in the adult mouse brain annotated by the Allen Brain Atlas (66). This analysis revealed that among the brain-specific ncRNA, many are potentially expressed solely in neurons, with some even being predominantly present in either glutamatergic (*Snord53*, *mir-802*) or GABAergic (*mir-3107*) neurons (Fig. 6*H*). Among glia-specific ncRNA, we identified *Snord17* and *mir-700* in macrophages as well as *mir-193a* and *mir-6236* in astrocytes and oligodendrocytes.

Discussion

Small ncRNA plays an indispensable role in shaping cellular identity in health and disease by orchestrating vital cellular processes and altering the expression of protein-coding genes (12). Recent efforts in profiling of the most studied types of small ncRNA, miRNA, across cells and tissues demonstrated the existence of tissue- and cell type-specific short noncoding transcripts (16, 24, 25, 33). In this work, we show that this phenomenon extends beyond one ncRNA class and involves not only tissue-specific but also sex-specific ncRNA expression. The present resource demonstrates that each healthy mammalian tissue carries a unique noncoding signature, contributed by well-understood RNA types as well as by less studied ones.

By analyzing the expression of several classes of ncRNA we discovered that nearly ∼900 transcripts contribute to the unique noncoding tissue profile. Moreover, we identified that in addition to variable transcription levels and posttranscriptional modifications, noncoding tissue specificity is achieved through an unknown mechanism of selective RNA retention. While at this point we are unable to judge the functional significance of this phenomenon, we discovered that, even between healthy tissues, certain miRNA undergo so-called "arm switching"—a process previously thought to be strictly pathogenic in mammals (43, 45, 67). Among other ncRNA class, tRFs, we observed a selective enrichment of certain fragment types over others, happening in both gene- and tissue-specific manners. Taken together with previous observations (7, 68, 69), this finding raises additional questions regarding the biogenesis pathways of tRFs as well as their tissue-specific function.

Within our study, we also report several tissue-specific miRNAs not identified in previous studies (Fig. 3, *SI Appendix*, Fig. S4, and Dataset S7). The validation process of the identified miRNAs brought to light several important observations. First, we noted that the AGO-CLIP, while often used as a "gold standard" of miRNA validation (24), in fact, does not support the existence of many miRNA independently detected within RNA-seq datasets or directly validated through RT-qPCR. The gap between AGO-CLIP and small RNA-seq data in terms of data quality, diversity, and depth suggests that validating against AGO-CLIP data may not be the optimal approach for miRNA discovery. Instead, one could search for evidence of miRNA expression within publicly available RNA-seq data as a first thresholding step (70). Second, it is important to consider that the genomic location of a novel miRNA might match with that of a previously annotated one, while the molecule itself could be transcribed from an opposite DNA strand. We observed this phenomenon on the de novo identified miRNA 14_6588, whose coordinates strictly overlap with *mir-802a* and that is only present in the brain.

miRNA has been previously used to train classifiers capable of differentiating cancer/tissue types (71). Our work demonstrates that machine learning algorithms applied to quantitative miRNA expression estimates also detect changes related to the cell-type composition of tissues, such as the shift in hematopoietic cell abundance in the postnatal compared to the fetal liver. Given the emerging evidence of ncRNA stability in the blood and its rapid propagation throughout the body within extracellular vesicles (72), we anticipate that the current space of markers used to noninvasively monitor development (73) could be further expanded to small ncRNA.

Small ncRNAs have been long known to regulate the development and function of the brain. Despite the tremendous progress of neuroscience in understanding the regulation of coding genes, surprisingly little is known about cell type-specific small ncRNA in the brain. Even within the available tissue-level ncRNA resources, the brain remains one of the most underrepresented tissue. We believe this is mostly due to technical limitations of small RNA sequencing, which has yet to be applied to single neurons and, so far, still relies on the robust enrichment of certain cell types. Given the extensive molecular heterogeneity of cell types in the brain, one would expect the diversity of ncRNAs in this tissue to be high. Our study finds that the brain, in fact, contains the largest number of unique mammalian ncRNA transcripts that are absent in other tissues. However, our knowledge of cell-specific ncRNA expression is not complete, and thus for the majority of these identified RNA we could not call the cell type of origin based in the data generated within previous studies offering cell type resolution (24, 25). Taking an alternative route and integrating our tissue-level ncRNA measurement with single-cell chromatin accessibility profiles turned out to be surprisingly informative and allowed us to infer the activity of ncRNA within individual neuronal and glial types. While the validation of these cell-specific transcripts through a direct measurement remains highly desirable, the provided ncRNA estimates indicate that ncRNA is another contributor to complexity in the architecture of nervous system.

We found that the lung contains the largest number of distinct small ncRNA among 11 profiled tissues. However, in the case of the lung, open chromatin data did not provide sufficient resolution for us to infer the cell types of origin for the majority of the transcripts. This inability to fully explain the roots of tissue complexity points to the need for further characterization of the ncRNAs content of specific cell types or even, similarly to mRNA, that of single cells (17, 74). This atlas, meanwhile, will hopefully stimulate future small ncRNA studies and serve as a powerful resource of ncRNA tissue identity for fundamental and clinical research.

Materials and Methods

Subject Details.

Animals. All procedures followed animal care and biosafety guidelines approved by Stanford University's Administrative Panel on Laboratory Animal Care and Administrative Panel of Biosafety. Wild-type C57BL/6J mice, 4 males and 10 females, aged ∼3 mo old, were used (Dataset S1).

Tissue Handling and RNA Extraction. Upon collection, tissue samples were submerged and preserved at −80 °C in RNAlater stabilization solution (Thermo Fisher; catalog #AM7021) until further processing. Total RNA was isolated from ∼100 mg of tissue using Qiagen miRNeasy mini kit (catalog #217004) and the Qiagen tissue lyser using 5-mm stainless-steel beads. RNA integrity was assessed using Agilent Bioanalyzer using RNA 6000 pico kit (Agilent Technologies; catalog #5067-1513).

Library Preparation and Sequencing. Short RNA libraries were prepared following the Illumina TruSeq Small RNA Library Preparation kit (catalog #RS-200-0012, RS-200-0024, RS-200-0036, RS-200-0048) according to the manufacturer's protocol and size-selected using Pippin Prep 3% Agarose Gel Cassette (Safe Science) in a range from 135 to 250 bp. Samples were pooled in batches of 48 and sequenced using the Illumina NextSeq500 instrument in a single-read, 50- or 75-base mode.

Data Processing. Sequencing reads were demultiplexed by BaseSpace (Illumina). Reads were trimmed from the adaptor sequences and aligned to the mouse genome (GRCm38) following ENCODE small RNA-seq pipeline (63), with minor modifications. We used STAR v2.5.1 (75) with the following parameters: –outFilterMismatchNoverLmax 0.04–outFilterMatchNmin 16–outFilterMatchNminOverLread 0–outFilterScoreMinOverLread 0–alignIntronMax 1–outMultimapperOrder Random–clip3pAdapterSeq TGGAATTCTC–clip3pAdapterMMp 0.1. We allowed incremental mismatch: no mismatches in the reads $\leq$25 bases, 1 mismatch in 26 to 50 bases, and 2 in 51 to 75 bases. Spliced alignment was disabled. We additionally filtered out reads "soft-clipped" at the 5′-end but kept 3′-clipped ones to account for miRNA isoforms and tRNA modifications. We used GENCODE M20 (29) and miRBase v22 (31) annotations to count the number of ncRNA transcripts. For snoRNAs, snRNAs, scaRNA, miscRNA, or miRNA quantification, reads were assigned to the respective genes using *featureCounts v 1.6.1* (76) with the following parameters: -a Mus_musculus.M20.gtf -M –primary -s 1. Read spanning two overlapping exons were excluded. To account for the multimappers, we used -M -primary option, which only counts a "primary" alignment reported by STAR (either a location with the best mapping score or, in the case of equal multimapping score, the genomic location randomly chosen as "primary"). This quantification approach largely agreed with the results obtained through mapping and quantification against the short nucleotide library (77) (*SI Appendix*, Fig. S9). However, it proved to be more inclusive for the reads uniquely mapping within the miRNA exon but missing one base at the 5′ prime end of the molecule and more strict in counting reads mapping elsewhere in the genome, for which the levels were consistently overestimated by the other method. All reads mapping to miRNA arms and to stem loops were used to quantify miRNA expression at the gene level. For tRF quantification, for each library, we first extracted reads mapped by STAR to the GENCODE-annotated tRNA within the mouse genome (30, 78). We then ran *unitas* (78) on these reads and used fractionated scores to compute the differences in tRF abundance across tissues.

Unsupervised Clustering and Dimensionality Reduction Analysis. Raw counts were normalized and log-transformed using *DESeq2* package. Batch effects were corrected using *limma* R package (79). Hierarchical clustering was performed using log$_2$-transformed expression values and using complete linkage as distance measure between clusters. We computed Euclidian distances between samples and used these values to perform the *t*-SNE with the following parameters: perplexity = 20 and maximum iteration of 1,000. Transcripts detected in one or more samples with overall log$_2$ expression scores <1 were excluded from this analysis.

TSI. To compute the tissue specificity index, we used the formula described previously in ref. 33:

$$TSI_j = \frac{\sum_{i=1}^{N}(1 - x_{j,i})}{N - 1},$$

where N is the total number of tissues measured and $x_{j,i}$ is the expression score of tissue i normalized by the maximal expression of any tissue for miRNA j.

Comparison with Available miRNA Data. To compute Spearman coefficients of correlation between samples generated in the current study and the mouse miRNA data generated by FANTOM5 consortium (25), we used *DESeq2*-normalized scores of 2,207 annotated miRNAs. To compare miRNA expression between mouse tissues and human cell types, we generated a curated list of miRNA orthologs, each of which contained a maximum of two mismatches per mature miRNA. In total, 531 miRNAs passed this criteria and were used to compute Spearman correlation coefficients shown in *SI Appendix*, Fig. S8.

Differential Expression Analysis with DESeq2. We used a likelihood-ratio test (LRT) implemented in *DESeq2* (80) to compute the significance of each gene in tissue-specific expression. Briefly, LRT compares whether the tissue type parameter, removed in the "reduced" (~ Batch + Sex), compared to "full" model (~ Batch + Sex +Tissue, in DESeq2), explains a significant amount of variation in the data. Statistical significance of the test (P values) was calculated by comparing the difference in deviance between the "full" and "reduced" model to χ^2 distribution. P values obtained from the LRT test were corrected using Benjamini–Hochberg procedure to obtain an FDR estimate of tissue specificity scores for each gene. Gene clusters in *SI Appendix*, Fig. S3*B* were computed on 250 differentially expressed genes ($P_{\text{adj}} <$ 1e-90 and base mean > 3) using *DEGreport* R package (81). miRNAs differentially expressed between female and male tissues were computed based on uniquely mapping counts (excluding multimappers), using Wald test within *DESeq2*. To test for the NULL hypothesis, we performed a permutation test in which we randomly reassigned the sex labels to 14 samples across each tissue and plotted the distribution of *DESeq2 P* values computed for the two groups (i.e., female and male) (*SI Appendix*, Fig. S7*A*). We used Benjamini–Hochberg-corrected P values (FDR) to assess the statistical significance of the computed DE scores (Fig. 5 and *SI Appendix*, Fig. S7*A*). The differentially expressed miRNAs were visualized on volcano plots, where male- and female-specific miRNAs (adjusted P value < 0.01 and absolute fold change > 1) were labeled accordingly.

Analysis of Correlation between miRNA Expression and the Expression of Its Targets. Putative miRNA target genes were extracted from TargetScan, DIANA, miRanda, or mirDB databases (82, 83). Only targets present in two or more databases were used. The gene expression scores of the respective targets in various tissues were extracted from the ENCODE database (84) (Dataset S7). Spearman correlation coefficients were computed between fragments per kilobase of transcript per million mapped reads retrieved from the ENCODE mRNA expression tables and *DESeq2*-normalized miRNA counts across 10 profiled tissues using corr.test() function from "psych" R package, and thresholded above Benjamini–Hochberg-adjusted P value of 0.1 and Spearman correlation coefficient ($-0.8 < r_s < 0.8$).

Identification of Identified Candidate miRNA. Candidate miRNAs were identified using miRDeep2 software (47). Only miRNAs supported by >5 reads were reported in this study. AGO-CLIP data were mapped to the mouse genome using STAR (same as for small RNA-seq libraries described in *Data Processing*) and the reads falling within the putative miRNA coordinates were counted using *featureCounts*. We counted a putative miRNA as "supported" if it had >5 AGO-CLIP counts.

To search for the previous mentions of identified miRNAs, we looked up their sequences in miRCarta (85) and used the Google search engine to query the literature. Candidate miRNAs were ranked by novoMiRank scores, which we computed as described in ref. 85.

For independent validation, we performed RT-qPCR using custom Small RNA TaqMan probes (Life Technologies; catalog #4398987) designed on the star consensus sequence reported by miRDeep2. We used 0.5 ng of total RNA per tissue sample supplied with Cel-mir-39 spike-in (Qiagen; catalog #339390) to perform the reactions in a final volume of 20 μL.

We analyzed tissue and sex specificity of identified miRNAs based on transcripts supported by at least 50 sequencing reads across all samples. Statistical analysis and data visualization were performed as described above for annotated miRNAs.

miRNA-Based Classifier. We trained the radial kernel SVM model on 136 samples corresponding to different tissue types (*SI Appendix*, Fig. S8*A*) using *e1070* (86) R package. We used *z* scores of *DESeq2* normalized counts obtained in this study as the train dataset and those obtained from ENCODE miRNA-seq data as the test dataset (Dataset S6). We normalized and scaled train and test datasets separately.

To measure the predictive power of each model we used the accuracy measure, calculated as the following:

$$\frac{\Sigma\,\text{True positive} + \Sigma\,\text{True negative}}{\Sigma\,\text{True observations}}.$$

We tuned the SVM model to derive optimal cost and gamma using tune.svm() function and searching within gamma ∈ [2^(−10): 2^10] and cost ∈ [10^(−5):10^3]. We tuned RF model using first random and then grid search, with an evaluation metric set to "Accuracy." The accuracy was computed using 10-fold cross-validation procedure. The reported accuracy is computed as a mean over the 10 testing sets in which nine folds are used for training and the held-out fold used as a test set. The R script used to train the models and compute the predictions is included in the supplement.

Comparison with scATAC-Seq Data. To compute and plot the correlations of small RNA-seq with scATAC-seq (Fig. 6 *B*–*G*), we used Cicero "activity scores" reported in Cusanovich et al. (28). Cicero scores were computed as described in ref. 87. Briefly, Cicero activity score represents the summarized score of chromatin accessibility of all sites linked to a given gene, which include proximal sites to the gene's transcription start site (within 500 bp of an annotated TSS) or distal sites linked to them. Cicero scores were loaded in *Seurat v3,* normalized, scaled, and averaged per cell or tissue type. To compute the accessibility scores for the brain-specific ncRNA in Fig. 6*H*, we used Cicero to derive gene activity scores from scATAC-seq data generated by 10XGenomics for the mouse adult brain (https://www.10xgenomics.com/10x-university/single-cell-atac/) with the chromium single-cell ATAC platform, and demultiplexed and preprocessed with the single-cell ATAC Cell Ranger platform. Using *Seurat v3*, we clustered the cells and merged them with Allen Brain Atlas single-cell RNA-seq data (66) for the further transfer of cell annotation labels. We computed the activity scores for brain-specific ncRNA identified through small RNA-seq using *cicero* (87). We used Spearman correlation of top 400 tissue-specific genes to compute the relationship between small RNA-seq and ATAC-seq activity scores reported in Fig. 6*B*.

Data Availability. The datasets generated and analyzed in the study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) repository (GSE119661) (88). All study data are included in the article and *SI Appendix*.

ACKNOWLEDGMENTS. We thank Dylan Henderson for assistance in RNA extraction and library preparation, Norma Neff and Jennifer Okamoto for sequencing expertise, Geoff Stanley and Kiran Kocherlakota for kind advice in tissue dissection and preservation, and Jennifer Okamoto and Norma Neff for the assistance in sequencing of the small RNA-seq libraries. This study was supported by Chan Zuckerberg Biohub. A.I. was supported by the Swiss National Foundation Early PostDoc Mobility Fellowship.

1. D. P. Bartel, Metazoan microRNAs. *Cell* **173**, 20–51 (2018).
2. T. R. Cech, J. A. Steitz, The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).
3. M. Esteller, Non-coding RNAs in human disease. *Nat. Rev. Genet.* **12**, 861–874 (2011).
4. J. Gebetsberger, L. Wyss, A. M. Mleczko, J. Reuther, N. Polacek, A tRNA-derived fragment competes with mRNA for ribosome binding and regulates translation during stress. *RNA Biol.* **14**, 1364–1373 (2017).
5. D. Becker *et al*., Nuclear pre-snRNA export is an essential quality assurance mechanism for functional spliceosomes. *Cell Rep.* **27**, 3199–3214.e3 (2019).
6. L. F. R. Gebert, I. J. MacRae, Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.* **20**, 21–37 (2019).
7. A. G. Telonis *et al*., Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget* **6**, 24797–24822 (2015).
8. S. L. Reichow, T. Hamma, A. R. Ferré-D'Amaré, G. Varani, The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res.* **35**, 1452–1464 (2007).
9. A. Keller *et al*., Toward the blood-borne miRNome of human diseases. *Nat. Methods* **8**, 841–843 (2011).

10. M. Somel *et al*., MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res.* **20**, 1207–1218 (2010).
11. M. Ha, V. N. Kim, Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–524 (2014).
12. A. G. Matera, R. M. Terns, M. P. Terns, Non-coding RNAs: Lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **8**, 209–220 (2007).
13. T. Kiss, Biogenesis of small nuclear RNPs. *J. Cell Sci.* **117**, 5949–5951 (2004).
14. H. Ishizu, H. Siomi, M. C. Siomi, Biology of PIWI-interacting RNAs: New insights into biogenesis and function inside and outside of germlines. *Genes Dev.* **26**, 2361–2373 (2012).
15. P. Kumar, J. Anaya, S. B. Mudunuri, A. Dutta, Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol.* **12**, 78 (2014).
16. P. Landgraf *et al*., A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).
17. O. R. Faridani *et al*., Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).
18. V. V. Sherstyuk *et al*., Genome-wide profiling and differential expression of microRNA in rat pluripotent stem cells. *Sci. Rep.* **7**, 2787 (2017).
19. S. Anfossi, A. Babayan, K. Pantel, G. A. Calin, Clinical utility of circulating non-coding RNAs—an update. *Nat. Rev. Clin. Oncol.* **15**, 541–563 (2018). .
20. F. J. Slack, A. M. Chinnaiyan, The role of non-coding RNAs in oncology. *Cell* **179**, 1033–1055 (2019).
21. H. L. A. Janssen *et al*., Treatment of HCV infection by targeting microRNA. *N. Engl. J. Med.* **368**, 1685–1694 (2013).
22. R. A. Ach, H. Wang, B. Curry, Measuring microRNAs: Comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. *BMC Biotechnol.* **8**, 69 (2008).
23. Y. Liang, D. Ridzon, L. Wong, C. Chen, Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics* **8**, 166 (2007).
24. M. N. McCall *et al*., Toward the human cellular microRNAome. *Genome Res.* **27**, 1769–1781 (2017).
25. D. de Rie *et al*.; FANTOM Consortium, An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* **35**, 872–878 (2017).
26. J. Jehn *et al*., 5′ tRNA halves are highly expressed in the primate hippocampus and sequence-specifically regulate gene expression. *RNA* **26**, 694–707 (2019).
27. J. M. Rimer *et al*., Long-range function of secreted small nucleolar RNAs that direct 2′-O-methylation. *J. Biol. Chem.* **293**, 13284–13296 (2018).
28. D. A. Cusanovich *et al*., A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018).
29. A. Frankish *et al*., GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
30. P. P. Chan, T. M. Lowe, GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189 (2016).
31. A. Kozomara, S. Griffiths-Jones, miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
32. V. Boivin *et al*., Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *RNA* **24**, 950–965 (2018).
33. N. Ludwig *et al*., Distribution of miRNA expression across human tissues. *Nucleic Acids Res.* **44**, 3865–3877 (2016).
34. R. C. Gallagher, B. Pils, M. Albalwi, U. Francke, Evidence for the role of PWCR1/HBII-85 C/D box small nucleolar RNAs in Prader–Willi syndrome. *Am. J. Hum. Genet.* **71**, 669–678 (2002).
35. R. A. Brady, V. M. Bruno, D. L. Burns, RNA-seq analysis of the host response to *Staphylococcus aureus* skin and soft tissue infection in a mouse model. *PLoS One* **10**, e0124877 (2015).
36. P.-Y. Chen *et al*., FGF regulates TGF-β signaling and endothelial-to-mesenchymal transition via control of let-7 miRNA expression. *Cell Rep.* **2**, 1684–1696 (2012).
37. A. Emde *et al*., Dysregulated miRNA biogenesis downstream of cellular stress and ALS-causing mutations: A new mechanism for ALS. *EMBO J.* **34**, 2633–2651 (2015).
38. M. P. Kowalski, T. Krude, Functional roles of non-coding Y RNAs. *Int. J. Biochem. Cell Biol.* **66**, 20–29 (2015).
39. S. J. Morrison, K. R. Prowse, P. Ho, I. L. Weissman, Telomerase activity in hematopoietic cells is associated with self-renewal potential. *Immunity* **5**, 207–216 (1996).
40. H. Liu, Y. Yang, Y. Ge, J. Liu, Y. Zhao, TERC promotes cellular inflammatory response independent of telomerase. *Nucleic Acids Res.* **47**, 8084–8095 (2019).
41. A. G. Matera, Z. Wang, A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15**, 108–121 (2014).
42. L. Pipes *et al*., The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res.* **41**, D906–D914 (2013).
43. L. Chen *et al*., miRNA arm switching identifies novel tumour biomarkers. *EBioMedicine* **38**, 37–46 (2018).
44. K. Pinel, L. A. Diver, K. White, R. A. McDonald, A. H. Baker, Substantial dysregulation of miRNA passenger strands underlies the vascular response to injury. *Cells* **8**, 83 (2019).
45. F. Kern *et al*., miRSwitch: Detecting microRNA arm shift and switch events. *Nucleic Acids Res.* **48**, W268–W274 (2020).
46. J. Alles *et al*., An estimate of the total number of true human miRNAs. *Nucleic Acids Res.* **47**, 3353–3364 (2019).
47. M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, N. Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).
48. J. M. Dhahbi *et al*., Deep sequencing identifies circulating mouse miRNAs that are functionally implicated in manifestations of aging and responsive to calorie restriction. *Aging (Albany NY)* **5**, 130–141 (2013).
49. T. Fehlmann *et al*., A high-resolution map of the human small non-coding transcriptome. *Bioinformatics* **34**, 1621–1628 (2018).
50. R. Javed *et al*., miRNA transcriptome of hypertrophic skeletal muscle with overexpressed myostatin propeptide. *BioMed Res. Int.* **2014**, 328935 (2014).
51. R. P. R. Metpally *et al*., Comparison of analysis tools for miRNA high throughput sequencing using nerve crush as a model. *Front. Genet.* **4**, 20 (2013).
52. L. S. Sundaram, "*Toxoplasma gondii*-mediated host cell transcriptional changes lead to metabolic alterations akin to the Warburg effect," PhD thesis, University of Cambridge, Cambridge, UK (2017).
53. Y. S. Lee, Y. Shibata, A. Malhotra, A. Dutta, A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* **23**, 2639–2649 (2009).
54. S. P. Thomas, T. T. Hoang, V. T. Ressler, R. T. Raines, Human angiogenin is a potent cytotoxin in the absence of ribonuclease inhibitor. *RNA* **24**, 1018–1027 (2018).
55. R. A. Forman *et al*., The goblet cell is the cellular source of the anti-microbial angiogenin 4 in the large intestine post *Trichuris muris* infection. *PLoS One* **7**, e42248 (2012).
56. L. V. Hooper, T. S. Stappenbeck, C. V. Hong, J. I. Gordon, Angiogenins: A new class of microbicidal proteins involved in innate immunity. *Nat. Immunol.* **4**, 269–273 (2003).
57. L. Guo, Q. Zhang, X. Ma, J. Wang, T. Liang, miRNA and mRNA expression analysis reveals potential sex-biased miRNA expression. *Sci. Rep.* **7**, 39812 (2017).
58. P. Loher, E. R. Londin, I. Rigoutsos, IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. *Oncotarget* **5**, 8790–8802 (2014).
59. C. M. Klinge, Estrogen regulation of microRNA expression. *Curr. Genomics* **10**, 169–183 (2009).
60. C. E. Fletcher, D. A. Dart, C. L. Bevan, Interplay between steroid signalling and microRNAs: Implications for hormone-dependent cancers. *Endocr. Relat. Cancer* **21**, R409–R429 (2014).
61. H. Guo, N. T. Ingolia, J. S. Weissman, D. P. Bartel, Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010).
62. C. Cortes, V. Vapnik, Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
63. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
64. M. H. Baron, J. Isern, S. T. Fraser, The embryonic origins of erythropoiesis in mammals. *Blood* **119**, 4828–4837 (2012).
65. U. Bissels, A. Bosio, W. Wagner, MicroRNAs are shaping the hematopoietic landscape. *Haematologica* **97**, 160–167 (2012).
66. E. S. Lein *et al*., Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
67. M. Lin *et al*., Comprehensive identification of microRNA arm selection preference in lung cancer: miR-324-5p and -3p serve oncogenic functions in lung cancer. *Oncol. Lett.* **15**, 9818–9826 (2018).
68. J. M. Dhahbi *et al*., 5′ tRNA halves are present as abundant complexes in serum, concentrated in blood cells, and modulated by aging and calorie restriction. *BMC Genomics* **14**, 298 (2013).
69. U. Sharma *et al*., Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science* **351**, 391–396 (2016).
70. C. Backes *et al*., Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.* **44**, e53 (2016).
71. M. Sherafatian, Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene* **677**, 111–118 (2018).
72. R. Bhome *et al*., Exosomal microRNAs (exomiRs): Small molecules with a big role in cancer. *Cancer Lett.* **420**, 228–235 (2018).
73. T. T. M. Ngo *et al*., Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360**, 1133–1136 (2018).
74. C. Trapnell, Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
75. A. Dobin *et al*., STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
76. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
77. Y. Lu, A. S. Baras, M. K. Halushka, miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics* **19**, 275 (2018).
78. D. Gebert, C. Hewel, D. Rosenkranz, unitas: The universal tool for annotation of small RNAs. *BMC Genomics* **18**, 644 (2017).
79. M. E. Ritchie *et al*., Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
80. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
81. L. Pantano, DEGreport: Report of DEG analysis (R package, Version 1.18.1). lpantano.github.io/DEGreport/. Accessed 1 January 2020.
82. V. Agarwal, G. W. Bell, J.-W. Nam, D. P. Bartel, Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
83. M. D. Paraskevopoulou *et al*., DIANA-microT web server v5.0: Service integration into miRNA functional analysis workflows. *Nucleic Acids Res.* **41**, W169–W173 (2013).
84. E. Pennisi, Genomics. ENCODE project writes eulogy for junk DNA. *Science* **337**, 1159–1161 (2012).
85. C. Backes *et al*., miRCarta: A central repository for collecting miRNA candidates. *Nucleic Acids Res.* **46**, D160–D167 (2018).
86. D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc functions of the Department of Statistics, Probability Theory Group (formerly: E1071), TU Wien (2017). https://cran.r-project.org/web/packages/e1071/e1071.pdf. Accessed 1 January 2020.
87. H. A. Pliner *et al*., Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).
88. A. Isakova, S. Quake, A mouse tissue atlas of small noncoding RNA. GEO (Gene Expression Omnibus). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119661. Deposited 7 September 2018.

Fehlmann *et al. Clinical Epigenetics* (2016) 8:123
DOI 10.1186/s13148-016-0287-1

Clinical Epigenetics

RESEARCH Open Access

cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs

Tobias Fehlmann[1], Stefanie Reinheimer[3], Chunyu Geng[2*], Xiaoshan Su[2], Snezana Drmanac[2,4], Andrei Alexeev[2,4], Chunyan Zhang[2], Christina Backes[1], Nicole Ludwig[3], Martin Hart[3], Dan An[2], Zhenzhen Zhu[2], Chongjun Xu[2,4], Ao Chen[2], Ming Ni[2], Jian Liu[2], Yuxiang Li[2], Matthew Poulter[2], Yongping Li[2], Cord Stähler[1], Radoje Drmanac[2,4], Xun Xu[2*], Eckart Meese[3] and Andreas Keller[1*]

Abstract

Background: We present the first sequencing data using the combinatorial probe-anchor synthesis (cPAS)-based *BGISEQ-500* sequencer. Applying cPAS, we investigated the repertoire of human small non-coding RNAs and compared it to other techniques.

Results: Starting with repeated measurements of different specimens including solid tissues (brain and heart) and blood, we generated a median of 30.1 million reads per sample. 24.1 million mapped to the human genome and 23.3 million to the *miRBase*. Among six technical replicates of brain samples, we observed a median correlation of 0.98. Comparing BGISEQ-500 to HiSeq, we calculated a correlation of 0.75. The comparability to microarrays was similar for both BGISEQ-500 and HiSeq with the first one showing a correlation of 0.58 and the latter one correlation of 0.6. As for a potential bias in the detected expression distribution in blood cells, 98.6% of HiSeq reads versus 93.1% of BGISEQ-500 reads match to the 10 miRNAs with highest read count. After using miRDeep2 and employing stringent selection criteria for predicting new miRNAs, we detected 74 high-likely candidates in the cPAS sequencing reads prevalent in solid tissues and 36 candidates prevalent in blood.

Conclusions: While there is apparently no ideal platform for all challenges of miRNome analyses, cPAS shows high technical reproducibility and supplements the hitherto available platforms.

Keywords: Next-generation sequencing, miRNA, Biomarker discovery, BGISEQ

Background

Currently, high-throughput analytical techniques are massively applied to further the understanding of the non-coding transcriptome [1]. Still, the full complexity of non-coding RNAs is only partially understood. One class of well-studied non-coding RNAs comprises small oligonucleotides, so-called miRNAs [2, 3].

Among the techniques most commonly used for miRNA profiling are microarrays, RT-qPCR, and next-generation sequencing (NGS), also referred to as high-throughput sequencing (HTS). An excellent review on the different platforms and a cross-platform comparison has been recently published [4]. A detailed examination of technologies, however, frequently reveals a bias. One reason for the respective bias is the ligation step, as, e.g., reported by Hafner and co-workers [5]. For example, the quantification of miRNAs differs between NGS and microarrays as it is dependent on base composition [6]. Especially, the guanine and uracil content of a miRNA seems to influence the abundance depending on the platform used. A substantial strength of NGS is the ability to support the completion of the non-coding transcriptome. Unlike microarrays and RT-qPCR, NGS allows the discovery of novel miRNA candidates. To this end, different algorithms have been implemented, with *miRDeep* being one of the most popular ones [7]. A substantial part of small RNA sequencing data has been obtained using HiSeq and MiSeq platforms (Illumina) based on stepwise sequencing by polymerase on DNA microarrays prepared by bridge PCR [8], as well as the

* Correspondence: gengchunyu@genomics.cn; xuxun@genomics.cn; andreas.keller@ccb.uni-saarland.de
[2]BGI-Shenzhen, Shenzhen, China
[1]Clinical Bioinformatics, Saarland University, 66125 Saarbrücken, Germany
Full list of author information is available at the end of the article

IonTorrent systems from Thermo Fisher Scientific using a different type of polymerase-based stepwise sequencing on micro-bead arrays generated by emulsion PCR, the first method proposed for making microarrays for massively parallel sequencing [9]. Another approach is the ligase-based stepwise sequencing also using micro-bead arrays, applied for example by ThermoFisher Scientific's SOLiD sequencing platform, and which has also been used to analyze and present novel miRNAs [10].

In the current study, we applied the new combinatorial probe-anchor synthesis (cPAS)-based BGISEQ-500 sequencing platform that combines DNA nanoball (DNB) nanoarrays [11] with stepwise sequencing using polymerase. An important advantage of this technique compared to the previously mentioned sequencing systems is in that no PCR is applied in preparing sequencing arrays. Applying cPAS, we investigated the human non-coding transcriptome. We first evaluated the reproducibility of sequencing on standardized brain and heart samples, then compared the performance to Agilent's microarray technique and finally evaluated blood samples. Using the web-based miRNA analysis pipeline *miRmaster* and the tool *novoMiRank* [12], we finally predicted 135 new high-likely miRNA candidates specific for tissue and 35 new miRNA candidates specific for blood samples.

Methods

Samples

In this study, we examined the performance of three sample types using three techniques for high-throughput miRNA measurements (Illumina's HiSeq sequencer, Agilent's miRBase microarrays, and BGI's BGISEQ-500 sequencing system, see details below). The three specimens were standardized HBRR sample ordered from Ambion (catalog number AM6051) and UHRR sample ordered from Agilent (catalog number 740000). UHRR and HBRR samples were measured in two and six replicates, respectively. As third sample type, we used *PAXGene* blood tubes. Here, two healthy volunteers' blood samples were collected and miRNAs were extracted using PAXgene Blood RNA Kit (Qiagen) according to manufacturer's protocol. The study has been approved by the local ethics committee.

Next-generation sequencing using BGISEQ-500

We prepared the libraries starting with 1 μg total RNA for each sample. Firstly, we isolated the microRNAs (miRNA) by 15% urea-PAGE gel electrophoresis and cut the gel from 18 to 30 nt, which corresponds to mature miRNAs and other regulatory small RNA molecules. After gel purification, we ligated the adenylated 3′ adapter to the miRNA fragment. Secondly, we used the RT primer with barcode to anneal the 3′ adenylated adapter in order to combine the redundant unligated 3′ adenylated adapter. Then, we ligated the 5′ adapter and did reverse transcript (RT) reaction. After cDNA first strand synthesis, we amplified the product by 15 cycles. We then carried out the second size selection operation and selected 103–115 bp fragments from the gel. This step was conducted in order to purify the PCR product and remove any nonspecific products. After gel purification, we quantified the PCR yield by Qubit (Invitrogen, Cat No. Q33216) and pooled samples together to make a single strand DNA circle (ssDNA circle), which gave the final miRNA library.

DNA nanoballs (DNBs) were generated with the ssDNA circle by rolling circle replication (RCR) to enlarge the fluorescent signals at the sequencing process as previously described [11]. The DNBs were loaded into the patterned nanoarrays and single-end read of 50 bp were read through on the BGISEQ-500 platform for the following data analysis study. For this step, the BGISEQ-500 platform combines the DNA nanoball-based nanoarrays [11] and stepwise sequencing using polymerase, as previously published [13–15]. The new modified sequencing approach provides several advantages, including among others high throughput and quality of patterned DNB nanoarrays prepared by linear DNA amplification (RCR) instead of random arrays by exponential amplification (PCR) as, e.g., used by Illumina's HiSeq and longer reads of polymerase-based cycle sequencing compared to the previously described combinatorial probe-anchor ligation (cPAL) chemistry on DNB nanorrays [11]. The usage of linear DNA amplification instead of exponential DNA amplification to make sequencing arrays results in lower error accumulation and sequencing bias.

Next-generation sequencing using HiSeq

Samples have been sequenced using Illumina HiSeq sequencing according to manufacturer's instructions and as previously described [16, 17].

Agilent microarray measurements

For detection of known miRNAs, we used the SurePrint G3 8×60k miRNA microarray (miRBase version 21, Agilent Technologies) containing probes for all miRNAs from miRBase version 21 in conjunction with the miRNA Complete Labeling and Hyb Kit (Cat. No. 5190-0456) according to the manufacturer's recommendations. In brief, 100 ng total RNA including miRNAs was dephosphorylated with calf intestine phosphatase. After denaturation, Cy3-pCp was ligated to all RNA fragments. Labeled RNA was then hybridized to an individual 8×60k miRNA microarray. After washing, array slides were scanned using the Agilent Microarray Scanner G2565BA with 3-μm resolution in double-pass mode. Signals were retrieved using Agilent AGW Feature Extraction software (version 10.10.11).

Data availability

The new sequencing data using BGISEQ-500 data are available in the Additional file of this manuscript (Additional file 1: Table S3).

Bioinformatics analysis

The raw reads were collapsed and used as input for the web-based tool miRMaster, allowing for integrated analysis of NGS miRNA data. On the server side, mapping to the human genome was carried out using *Bowtie* [18] (one mismatch allowed). miRNAs were quantified similar to the popular *miRDeep2* [19] algorithm. The prediction of novel miRNAs was performed using an extended feature set built up on novoMiRank [12]. For classification, an *AdaBoost* model using decision trees was applied. Novel miRNAs were cross-checked against other RNA resources, including the *miRBase* [20], *NONCODE2016* [21], and *Ensembl* non-coding RNAs. The assessment of the quality of new miRNAs was carried out using the novoMiRank algorithm. A downstream analysis of results including cluster analysis was performed using R. For target prediction, we applied TargetScan 7.1 (http://www.targetscan.org/vert_71/) and predicted for all new miRNAs the targets. With the predictions, we extracted the context ++ scores and used them for prioritizing the targets, miRNA-target interactions with context++ scores below 1 were considered as high-likelihood targets. Target networks were constructed using an offline version of MiRTargetLink [22] and visualized in Cytoscape. miRNA target pathway analysis has been carried out using GeneTrai2 [23]. For the GeneTrail2 analysis, all available categories were analyzed, the minimal category size was set to 4 and all *p* values were adjusted using Benjamini-Hochberg adjustment.

Results

Raw data analysis

We sequenced six brain, two heart, and two blood samples using the BGISEQ-500 system. The resulting reads were mapped to the human genome allowing one mismatch per read. The 10 samples had a median of 30.1 million reads. Of these, 24.1 million reads mapped to the human genome and 23.3 million reads to miRNAs annotated in the human miRBase version 21. The remaining 0.7 million reads per sample contain potentially new miRNAs.

Technical reproducibility of the BGISEQ-500 and comparison to microarrays

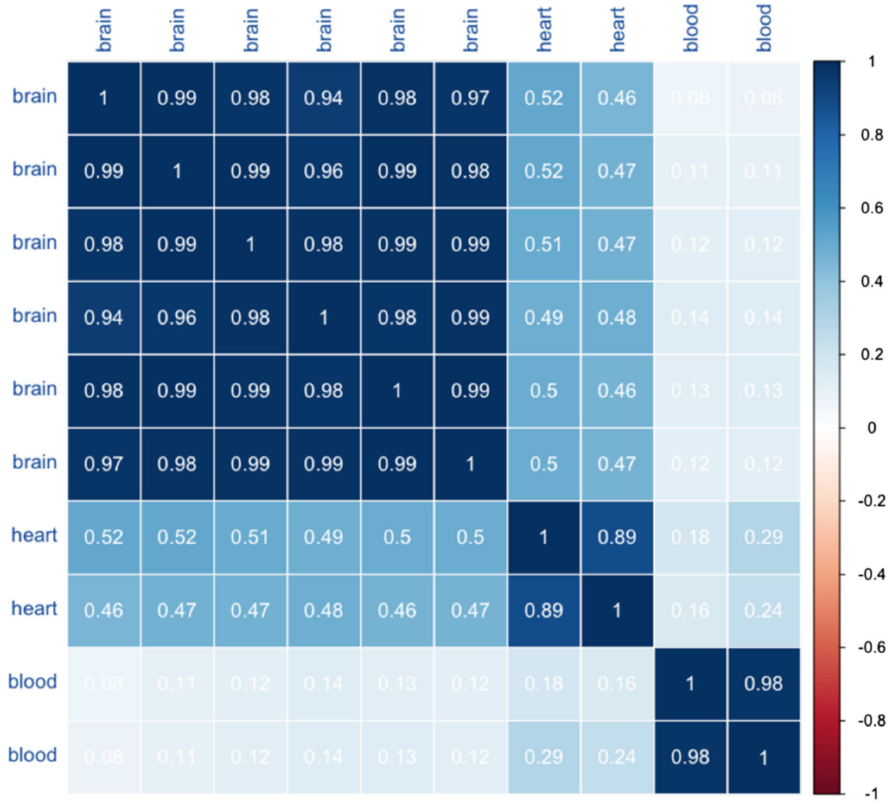
To assess the technical reproducibility of the sequencing platform, we evaluated the six technical replicates of the human brain sample (see correlation matrix in Fig. 1). The median correlation between the six replicates was 0.98, and the 25 and 75% quantile were 0.98 and 0.99, respectively. These data suggest an overall high correlation for technical replicates on the BGISEQ-500 platform.

Comparing the BGISEQ-500 data to the measurements of the brain sample with microarrays (miRBase version 21) that have also been carried out as six technical replicates (median correlation of the microarrays was 0.999), we observed a log correlation of 0.48. A direct comparison is presented in the scatter plot in Fig. 2a. This plot highlights many miRNAs that can be measured at a comparable level on both platforms. However, a subset of the small non-coding RNAs is shifted towards higher expression on the array platform. The same behavior can be observed in the cluster heat map in Fig. 2b. This heat map graphically represents the 50 miRNAs with most different detection between both techniques. To compare rather the ranks of miRNAs instead of the absolute read counts, the replicated brain samples on both platforms were jointly quantile normalized. Three miRNAs, in particular, showed highly significant deviations (multiple testing adjusted *p* values below 10^{-20}). Hsa-miR-8069 was almost not detected in the BGISEQ-500 but had 0.9 million normalized intensity counts on the array platform, hsa-miR-4454 had 51.6 normalized reads on the BGISEQ-500 versus 1.9 million normalized counts on the microarrays, and hsa-miR-7977 had 343.2 normalized reads on the BGISEQ-500 versus 1.3 million normalized counts on the microarrays. This means that the three miRNAs were orders of magnitudes more abundant on microarrays as compared to the sequencing system. The secondary structures of the three precursors are presented in Additional file 2: Figure S1. These results match well to previously published platform comparisons between NGS and microarrays [6]. Here, several miRNAs such as hsa-miR-941 (not detected in any array experiment, not detected in RT-qPCR, average read count of ~1000 reads using Illumina HiSeq sequencing) had expression levels differing several orders of magnitude between the miRBase microarrays and using HiSeq sequencing.

The full list of miRNAs with raw and adjusted *p* values in *t* test and Wilcoxon-Mann-Whitney test comparing BGISEQ-500 and microarrays is presented in Additional file 3: Table S1. Overall, the results are well in-line with those obtained between HiSeq NGS and the same microarray platform [6]. Reasons that explain differences between arrays and NGS include different sensitivity levels of the platforms, cross-hybridization of miRNAs with similar sequences on the microarrays or bias in library preparation. Further, effects of the normalization can lead to variations in miRNA quantification.

Biological replicates of blood samples and comparison to other platforms

One of the most promising applications in small RNA analysis is biomarker profiling in body fluids. We



	brain	brain	brain	brain	brain	brain	heart	heart	blood	blood
brain	1	0.99	0.98	0.94	0.98	0.97	0.52	0.46	0.08	0.08
brain	0.99	1	0.99	0.96	0.99	0.98	0.52	0.47	0.11	0.11
brain	0.98	0.99	1	0.98	0.99	0.99	0.51	0.47	0.12	0.12
brain	0.94	0.96	0.98	1	0.98	0.99	0.49	0.48	0.14	0.14
brain	0.98	0.99	0.99	0.98	1	0.99	0.5	0.46	0.13	0.13
brain	0.97	0.98	0.99	0.99	0.99	1	0.5	0.47	0.12	0.12
heart	0.52	0.52	0.51	0.49	0.5	0.5	1	0.89	0.18	0.29
heart	0.46	0.47	0.47	0.48	0.46	0.47	0.89	1	0.16	0.24
blood	0.08	0.11	0.12	0.14	0.13	0.12	0.18	0.16	1	0.98
blood	0.08	0.11	0.12	0.14	0.13	0.12	0.29	0.24	0.98	1

Fig. 1 Correlation matrix of the brain (six technical replicates), heart (two technical replicates), and blood (two biological replicates) sequenced by the BGISEQ-500 system

previously analyzed over 2000 blood samples on Agilent microarrays [17, 24, 25] and about 1000 samples using HiSeq sequencing [26, 27] and compared both platforms [6]. We correlated two newly sequenced blood samples using the BGISEQ-500 system to the data generated by HiSeq and Agilent microarrays. When interpreting the results, it is important to keep in mind that the microarrays and HiSeq data are from the same samples [6] while the newly sequenced blood drawings are from other individuals and thus biological but no technical replicates. To minimize a potential bias between the platforms with respect to different miRNA sets, we first reduced the marker set to the 2525 human miRNAs that were profiled on all platforms and next to the subset of 658 miRNAs that were discovered in all three platforms. For each, platform data were normalized using quantile normalization. Due to the wide dynamic range of miRNAs in blood samples, which is approximately 10^7, we present the three pairwise comparisons (BGISEQ-500 to microarrays, BGISEQ-500 to HiSeq, and HiSeq to microarrays) on a log scale. The scatter plots are presented in Fig. 3. The highest correlation was observed for BGISEQ-500 to Illumina (0.75, Fig. 3a). Even the correlation between microarrays and HiSeq was below this value (0.6, Fig. 3c). Especially since technical replicates have been measured for these platforms, the increased correlation of sequencing platforms is remarkable. The comparison of BGISEQ-500 and microarrays revealed correlation values in the same range as for the brain samples (0.58, Fig. 3b). The 3D scatter plot in Fig. 3d compares the expression of the three platforms directly to each other. The coloring of the miRNAs has been carried out with respect to the GC content.

Expression distribution of miRNAs

As mentioned, miRNA expression is highly variable and can scatter across many orders of magnitude. We thus compared the distribution of the sequencing reads in blood samples on the HiSeq to the BGISEQ-500. Blood samples, including blood cells (especially red blood cells) are known to be enriched for few miRNAs that are highly expressed. The diagram in Fig. 4 (panel A) highlights that 90.8% of all blood sequencing reads from the HiSeq match to one single miRNA: hsa-miR-486-5p. The second most abundant miRNA miR-92a-3p takes further 5.5%, and already the third most abundant marker miR-451a has below 1% of all reads. In sum, 98.6% of all reads match to the top 10 miRNAs. For the

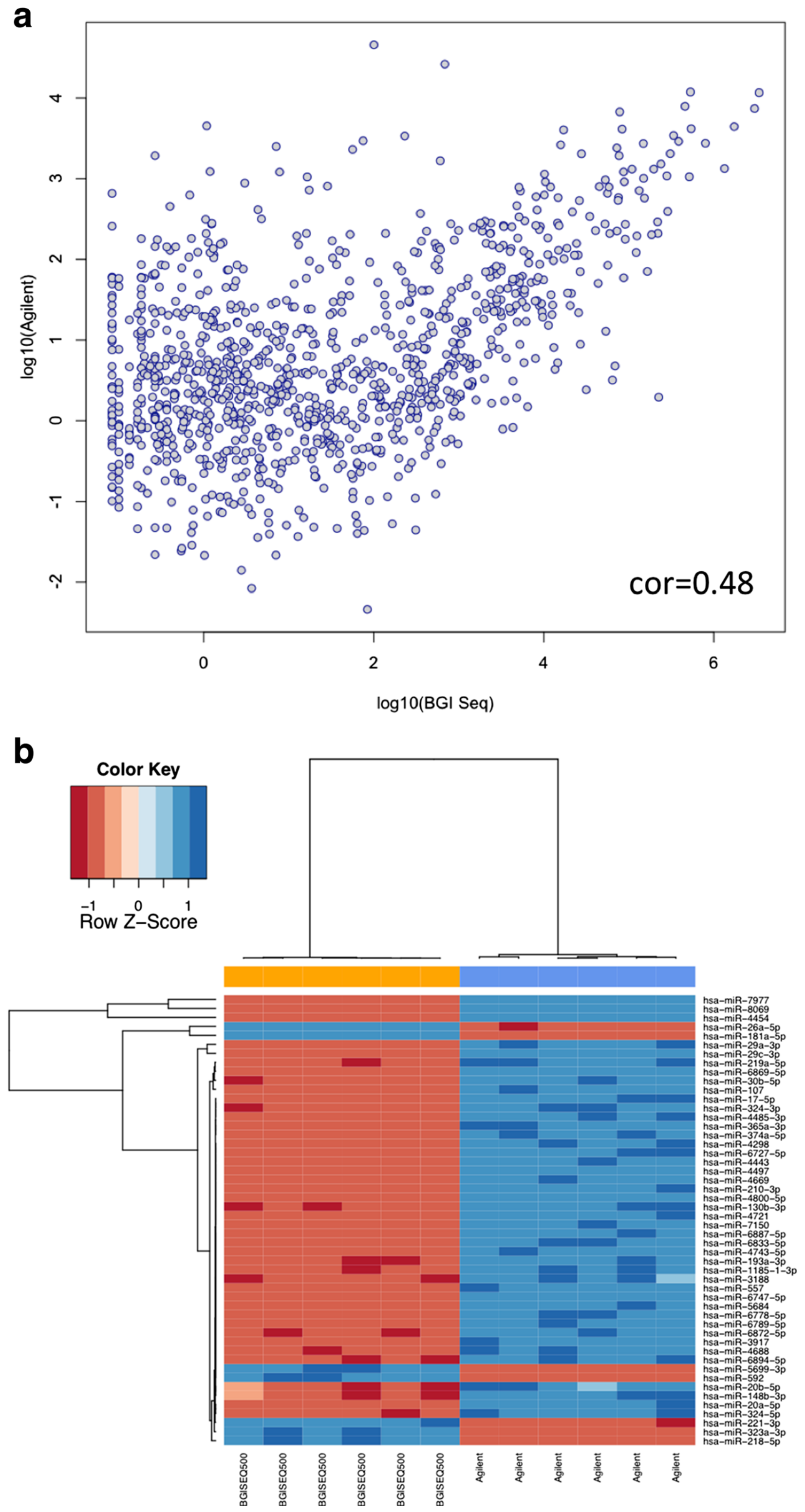


Fig. 2 a Log average expression of common miRNAs for the brain RNA on BGISEQ-500 and on Agilent microarrays (six technical replicates each). **b** Heat map with dendrogram for the 50 most differently detected miRNAs in the brain RNA between Agilent and BGISEQ-500 (six technical replicates each)

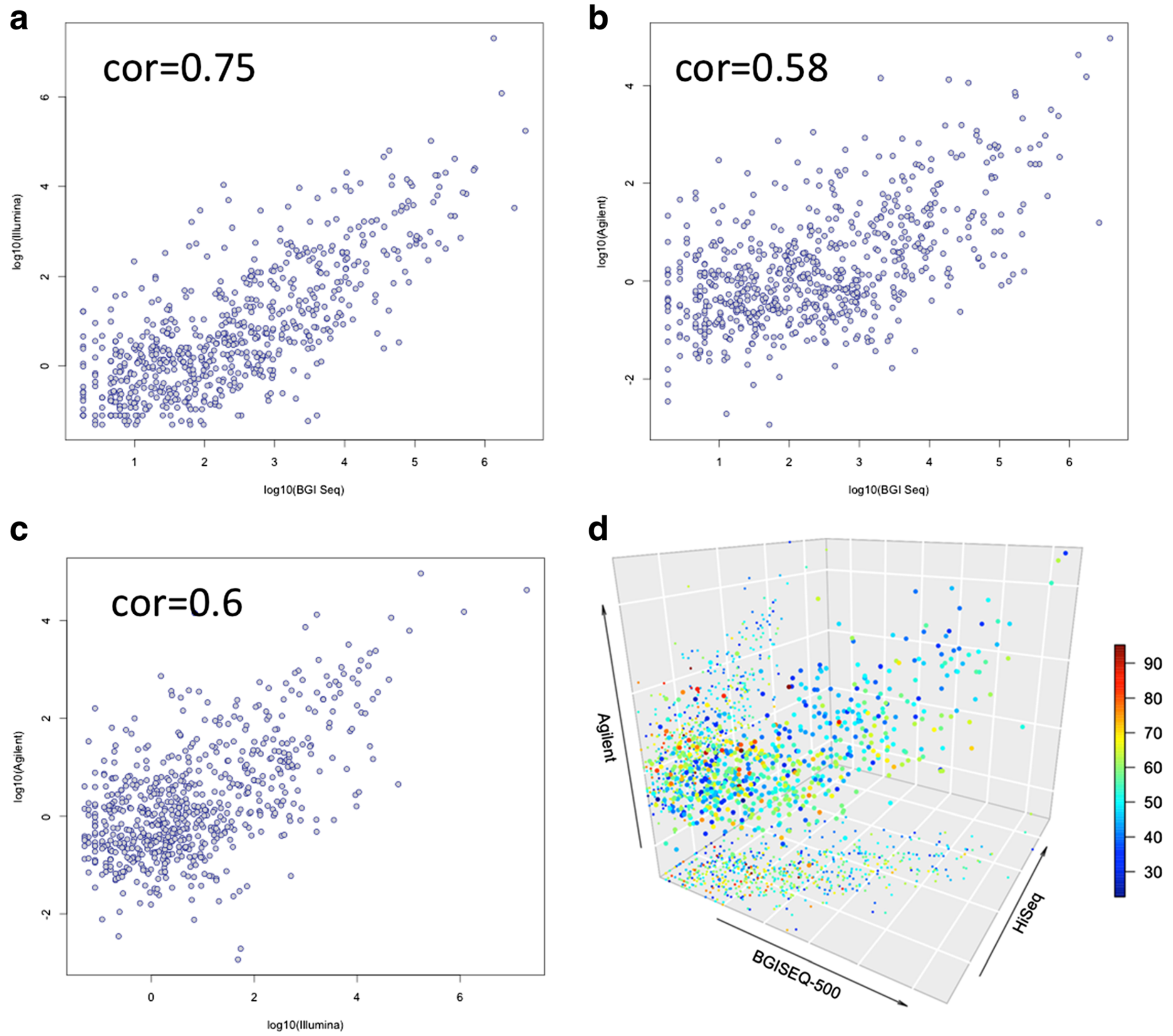


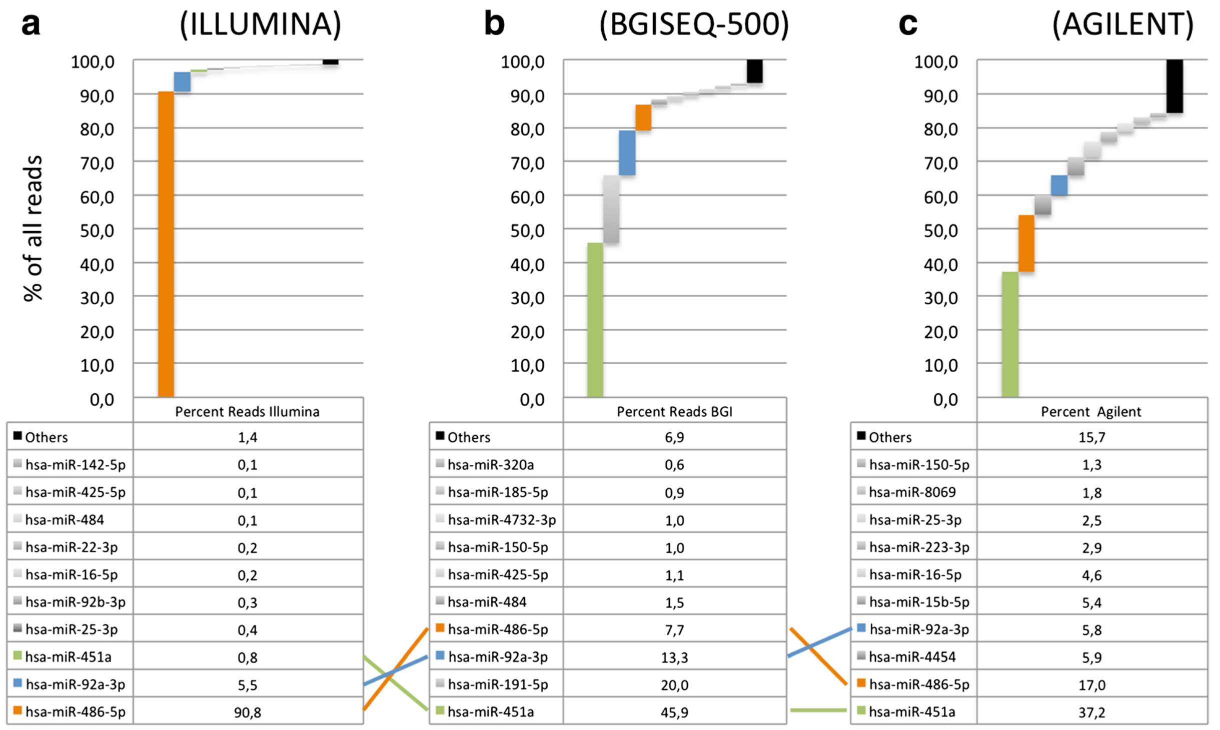
Fig. 3 **a-c** Pairwise scatter plots for comparing expression of miRNAs in blood cells on microarrays, HiSeq, and BGISEQ-500. Please note that for HighSeq and Agilent technical replicates were measured, for BGISEQ-500 biological replicates. **d** 3D scatter-plot colored by the GC content of miRNAs

BGISEQ-500 (panel B), 45.9% of reads match to miR-451a, further 20% map to miR-191-5p and 13.3% map to miR-92a-3p. The most abundant miRNA in HiSeq, miR-486-5p, is detected in 7.7% of all reads. 93.1% of all sequenced reads match to the top 10 miRNAs.

Comparison of the distribution and abundance of miRNAs on the microarray platform is difficult since microarrays show a saturation effect. This means that for two miRNAs expressed in a range above the saturation, no difference can be observed. We nonetheless performed the same analysis as presented above, assuming that the sum of all expression counts equals to 100%. In this analysis, miR-451a which is found in 0.8% of HiSeq reads and 45.9% of BGISEQ-500 reads is the highest expressed in microarrays (37.2% of all expression counts), followed by 17% of miR-486-5p.

Prediction of novel miRNAs

Predicting new miRNAs from NGS data is a challenging task since many false positive miRNA candidates are observed. We implemented our own prediction tool for miRNAs from NGS data and filtered the candidates stringently to reduce the false discovery rate. Without any filtering steps, our initial predictor trimmed for maximizing the ROC AUC returned 25,086 candidates across all samples. The exclusion of the candidates with low abundance (less than 10 total reads) reduced the number of candidates to around 10% (2354 candidates).



	Percent Reads Illumina
Others	1,4
hsa-miR-142-5p	0,1
hsa-miR-425-5p	0,1
hsa-miR-484	0,1
hsa-miR-22-3p	0,2
hsa-miR-16-5p	0,2
hsa-miR-92b-3p	0,3
hsa-miR-25-3p	0,4
hsa-miR-451a	0,8
hsa-miR-92a-3p	5,5
hsa-miR-486-5p	90,8

	Percent Reads BGI
Others	6,9
hsa-miR-320a	0,6
hsa-miR-185-5p	0,9
hsa-miR-4732-3p	1,0
hsa-miR-150-5p	1,0
hsa-miR-425-5p	1,1
hsa-miR-484	1,5
hsa-miR-486-5p	7,7
hsa-miR-92a-3p	13,3
hsa-miR-191-5p	20,0
hsa-miR-451a	45,9

	Percent Agilent
Others	15,7
hsa-miR-150-5p	1,3
hsa-miR-8069	1,8
hsa-miR-25-3p	2,5
hsa-miR-223-3p	2,9
hsa-miR-16-5p	4,6
hsa-miR-15b-5p	5,4
hsa-miR-92a-3p	5,8
hsa-miR-4454	5,9
hsa-miR-486-5p	17,0
hsa-miR-451a	37,2

Fig. 4 Expression distribution of the 10 miRNAs with the highest detection in the blood RNA on the HiSeq system (**a**), BGISEQ-500 (**b**), and microarray system (**c**). Note that for the Agilent microarray system, the sum of all expression intensities was assumed to be 100%

Further analysis with *novoMiRank* (cutoff 1.5) filtered out more miRNAs, leaving 1553. The miRNAs were flagged by *novoMiRank* because of a high deviation from miRNAs in the first *miRBase* versions, including deviating length, free energy, or nucleic acid composition of miRNAs. Matching the remaining candidates to other RNA resource in a blacklisting step finally presented 926 miRNA candidates (Additional file 4: Table S2). Still, it is likely that this set contains many false positives. Additionally, low-throughput experimental validation of almost 1000 miRNA candidates, e.g., by Northern Blot is a very labor-extensive approach. We thus additionally compared the frequency of reads mapping to the blood versus tissue samples. As detailed in Fig. 5a, we observe a substantial variability between blood and tissue for the 926 miRNA candidates (correlation 0.18). Defining a miRNA as tissue/blood specific if it occurs with a factor of 100-fold higher in one of both sample types (normalized for the total number of samples) highlighted 74 new miRNA candidates specific for tissue and 36 new miRNA candidates specific for blood samples. Figure 5b shows bar plots for two miRNA precursors, the most tissue specific novel-mir-36616 (blue), only present in the brain samples, and the blood specific novel-mir-31007. The first miRNA, which is observed exclusively in the brain samples and not in the heart, reveals a significantly less expressed 3′ mature form as compared to the 5′ mature form. The second miRNA is exclusively observed in blood samples. Here, the 5′ mature form is lower expressed compared to the 3′ form. The boxes above the bar plots show the secondary structures of both miRNA candidates.

miRNA target analysis

For all 926 miRNAs, we predicted targets using TargetScan. To rank miRNA-target interactions, we used the context++ score (distribution of the context++ score across all predictions is provided in Additional file 5: Figure S2). Thereby, we observed an accumulation of high-likelihood targets for tissue-specific miRNAs. Of the 926 miRNAs, the tissue specific had an average 42.8 targets, the neither for blood nor for tissue-specific miRNAs 40.7 targets while for blood-specific miRNAs, only 34.5 targets were predicted. The complex miRNA-target network is presented in Additional file 6: Figure S3. It contains 6014 nodes (5088 genes and 926 miRNAs). Network characteristics such as degree distribution and shortest path length are presented in Additional file 7: Figure S4. The genes with largest numbers of predicted miRNAs targeting the gene were CYB561D1 (229 miRNAs), FBXL12 (174 miRNAs), PML (162 miRNAs), and VNN3 (154 miRNAs). The distribution of miRNAs in

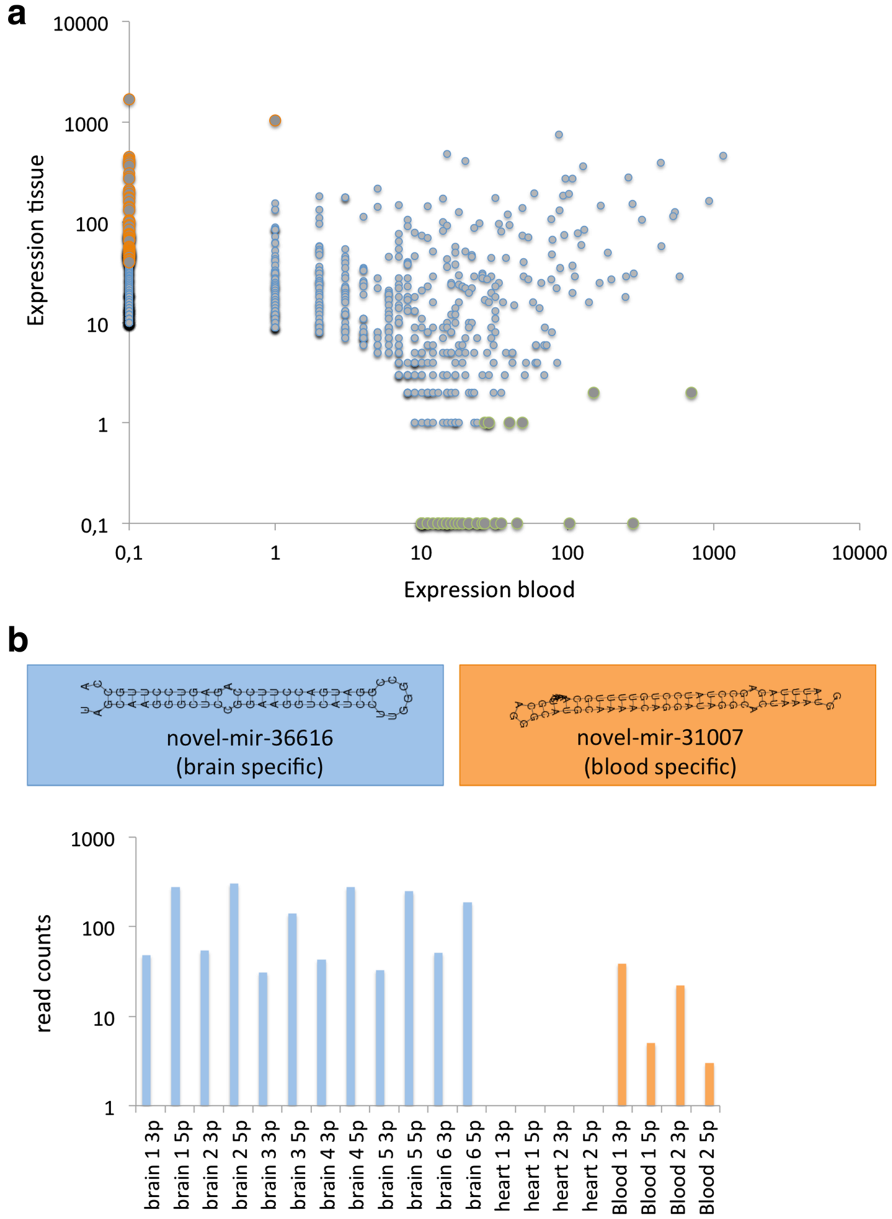


Fig. 5 a Expression of novel miRNAs in blood versus tissue. The *green* miRNAs are specific for blood, the *orange* miRNAs for tissue, and the *blue* miRNAs were detected in both specimens. **b** Bar plot for two different miRNAs. The first miRNA novel-mir-36616 in *blue* is detected only in the brain tissue and not in the heart tissue or blood, the -3p form is one order of magnitude less expressed compared to the -5p mature form. The second miRNA novel-mir-31007 is expressed in blood and not in tissue, here, the -3p form is more than an order of magnitude more abundant compared to the -5p form

the different group is presented as Venn diagram in Additional file 8: Figure S5). Among the predicted target genes that were found only for candidate miRNAs being blood specific was, e.g., HMOX1, heme oxygenase 1, mediating the first step of the heme catabolism by cleaving heme to build biliverdin or HPX, coding for hemopexin. The complex nature of the in silico calculated miRNA-target network requires further analyses to understand whether target genes accumulate in specific biochemical categories such as KEGG pathways or gene ontologies. We thus applied GeneTrail2 separately to the set of genes targeted by blood specific miRNAs, targeted by tissue specific miRNAs and by all other miRNAs. As the background sets, all genes predicted to be targeted by at least a single miRNA were selected and the functionality to compare different enrichment analyses by

GeneTrail2 has been used. Enriched pathways seem to be largely relevant for either blood or tissue miRNAs, as Additional file 9: Figure S6 highlights. Tissue specific miRNAs had target genes enriched for DNA damage response, the apoptosis, or RNA polymerase II regulatory region DNA binding while blood miRNAs target genes were, e.g., enriched for TP35 network. Interestingly, tissue miRNA target genes also clustered on specific genomic locations (e.g., 19p12 and 19.q13) while blood miRNA targets did not show such an enrichment. In contrast, blood miRNA targets were enriched for disease phenotypes such as carotid artery diseases. In sum, the enrichment analysis highlights very distinct patterns for blood and tissue miRNA targets. Of course, not only the new miRNAs themselves but also the predicted targets deserve detailed experimental validation.

Discussion

The advent of next-generation sequencing reduced the costs of sequencing while simultaneously increasing the speed of throughput [28]. Today, the costs for small RNA seq are almost equal to and even lower than miRNA microarrays, although small RNA-seq provides the additional possibility for detecting novel small RNA entities.

In the present study, we investigated two current sequencing approaches supporting massively parallel sequencing, which is of high relevance in small RNA research because of the high dynamic range of these molecules: DNA nanoball [11]-based sequencing by BGISEQ-500 and PCR cluster [8]-based sequencing by HiSeq. An important difference between these techniques is in that the first approach uses linear DNA amplification, and the second uses exponential DNA amplification to make sequencing arrays. The latter approach may in turn lead to amplification errors and some specific biases. Besides this fundamental difference, both approaches have their additional advantages and disadvantages. Specifically for the BGISEQ-500, the library preparation currently takes around three working days, the sequencing itself needs one or at maximum two working days. Each flowcell of the BGISEQ-500 has two lanes. On each of these lanes, 32 Gb data can be generated using single-end reads of length 50 bases. The cost of the reagent and material is around 200 USD for 20 million reads ensuring high-quality data at a reasonable cost.

Recently, we published a manuscript about bias in NGS and microarray analysis for miRNAs [6], highlighting that the expression of miRNAs on different platforms varies by, for example, the nucleic acid composition. In the validation by RT-qPCR, we focused on miRNAs discordant between the high-throughput platforms. Thereby, we observed cases where the RT-qPCR results were concordant with Illumina HiSeq, with microarrays or with none of the techniques. Therefore, we were especially interested how the BGISEQ-500 platform compares to the HiSeq platform and microarrays with the content from the *miRBase* for small RNA analysis.

Three miRNAs had high divergence between arrays and BGISEQ-500, among them hsa-miR-4454, which was high abundant in arrays but almost not detectable in BGISEQ-500. According to the miRBase, only 28% of users believe that this miRNA is real. Although such votes have only limited value, they at least indicate that this miRNA may be influenced by technological bias.

For high-throughput sequencing, the library preparation and the kits used play a crucial role for the quality of the sequencing results. Others and we noticed an overly abundance of the miRNA miR-486-5p when using the TruSeq kit (Illumina, San Diego), which seems to be independent of the source of the analyzed material [6, 29, 30]. Using the BGISEQ-500 platform, we observed lower read counts for this miRNA. However, in some cases, the miRNA abundance of BGISEQ-500 matches to the HiSeq sequencing results while microarrays show a different expression level, and in other cases, the BGISEQ-500 deviates from the other platforms and in several cases, all three techniques provide substantially divergent results. The more even distribution of reads of the BGISEQ-500 compared to the HiSeq results facilitates the discovery of new miRNAs, which are expected to be significantly less expressed as compared to the already known miRNAs, especially from early miRBase versions.

With respect to many miRNA currently annotated in miRBase and the rapidly growing number of new miRNAs, it is essential not only to have tools for filtering likely false-positives such as the NovoMiRank tool but also to carry out validation of miRNAs using other molecular biology approaches such as cloning and Northern blotting.

Focusing on the performance of the BGISEQ-500, we found a high technical reproducibility of sequencing results, which was however slightly below the technical reproducibility of microarrays. This fact can have different reasons, e.g., the different limit of detection of microarrays. In contrast to sequencing, microarrays have a saturation effect. With respect to the total number of discovered known miRNAs, performance of the BGISEQ-500 was comparable both to the Illumina and the microarray platform.

Conclusions

In sum, none of the mentioned platforms seems to provide the “ultimate solution” in miRNA analysis. All have their advantages and disadvantages and show some bias for the detection of certain sequence types.

Additional files

Additional file 1: Table S3. miRNA read count of the BGISEQ-500. (XLSX 250 kb)

Additional file 2: Figure S1. Predicted secondary structures for selected miRNAs. (PNG 241 kb)

Additional file 3: Table S1. Comparison of BGISEQ-500 to Agilent. (XLSX 135 kb)

Additional file 4: Table S2. List of novel miRNA candidates. (XLSX 6531 kb)

Additional file 5: Figure S2. Histogram of the decade logarithm of the context++ scores (multiplied by −1) of predicted targets for the candidate miRNAs. Since negative context++ scores are favorable, the miRNA targets on the right of the diagram are more likely true interactions. (PNG 78 kb)

Additional file 6: Figure S3. Full interaction network. Predicted miRNAs are represented in large nodes, colored by type (red: blood specific, blue: tissue specific, green: all others) and genes are represented by smaller gray nodes. (PNG 1033 kb)

Additional file 7: Figure S4. Core network characteristics as node degree distribution (*top*) and shortest path length (*bottom*). (PNG 129 kb)

Additional file 8: Figure S5. Venn diagram showing the distribution of predicted target genes for tissue-specific miRNA candidates, blood-specific miRNA candidates, and all other miRNA candidates. (PNG 156 kb)

Additional file 9: Figure S6. Comparison of the pathway enrichment analysis for the GeneTrail2 analysis with respect to the three target sets. *Red arrows* represent significant enrichments. (PNG 289 kb)

Acknowledgements
We acknowledge the support of BGI-Shenzhen and Complete Genomics.

Funding
The study has been funded by internal funds of Saarland University.

Availability of data and materials
Following publication expression data are available in the gene expression omnibus (GEO).

Authors' contributions
Setting up the assay were done by CG, XS, AA, SD, CZ, DA, JL, and RD. Generating miRNA data were done by SR, CZ, NL, MH, ZZ, CX, AC, and MN. Evaluation of data was done by TF, CB, NL, YL, and AK. Drafting and revision of the manuscript were done by EM, AK. Study design and set-up were done by YL, CS, XX, EM, and AK. All authors read and approved the final manuscript.

Competing interests
Authors with affiliations 1 and 2 are employed by BGI-Shenzhen, Shenzhen, China, and Complete Genomics (a BGI company), Mountain View, CA, USA.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The study has been approved by the local ethics committee (Ärztekammer des Saarlandes).

Author details
[1]Clinical Bioinformatics, Saarland University, 66125 Saarbrücken, Germany. [2]BGI-Shenzhen, Shenzhen, China. [3]Department of Human Genetics, Saarland University, Saarbrücken, Germany. [4]Complete Genomics (a BGI company), Mountain View, CA, USA.

Received: 6 October 2016 Accepted: 4 November 2016
Published online: 21 November 2016

References

1. Veneziano D, Nigita G, Ferro A. Computational approaches for the analysis of ncRNA through deep sequencing techniques. Front Bioeng Biotechnol. 2015;3:77.
2. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843–54.
3. Ruvkun G. Molecular biology. Glimpses of a tiny RNA world. Science. 2001;294(5543):797–9.
4. Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, D'Andrade P, DeMayo M, Dennis L, et al. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods. 2014;11(8):809–15.
5. Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 2011;17(9):1697–712.
6. Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, Meese E, Keller A. Bias in high-throughput analysis of miRNAs and implications for biomarker studies. Anal Chem. 2016;88(4):2088–95.
7. Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008;26(4):407–15.
8. Mayer P, Farinelli L, Kawashima EHUhwgcpUS. Method of nucleic acid amplification. In.: Google Patents; 2011
9. Drmanc R, Crkvenjakov R. Prospects for a miniaturized, simplified and frugal human genome project. Sci Yugosl. 1990;16(1–2):97–107.
10. Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, Vogel B, Matzas M, Huwer H, Katus HA, et al. Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. Mol BioSyst. 2011;7(12):3187–99.
11. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78–81.
12. Backes C, Meder B, Hart M, Ludwig N, Leidinger P, Vogel B, Galata V, Roth P, Menegatti J, Grasser F, et al. Prioritizing and selecting likely novel miRNAs from NGS data. Nucleic Acids Res. 2016;44(6):e53.
13. Canard B, Sarfati RS. DNA polymerase fluorescent substrates with reversible 3′-tags. Gene. 1994;148(1):1–6.
14. Tsien RY, Ross P, Fahnestock M, Johnston AJUhwgcpCAAce. Dna sequencing. In.: Google Patents; 1991
15. Church GM, Mitra RDUhwgcpEPAce. Nucleotide compounds having a cleavable linker. In.: Google Patents; 2003
16. Meder B, Backes C, Haas J, Leidinger P, Stahler C, Grossmann T, Vogel B, Frese K, Giannitsis E, Katus HA, et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. Clin Chem. 2014;60(9):1200–8.
17. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul F, Stahler C, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol. 2013;14(7):R78.
18. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
19. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40(1):37–52.
20. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34(Database issue):D140–4.
21. Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. 2016;44(D1):D203–8.
22. Hamberg M, Backes C, Fehlmann T, Hart M, Meder B, Meese E, Keller A. MiRTargetLink—miRNAs, genes and interaction networks. Int J Mol Sci. 2016;17(4):564.
23. Stockel D, Kehl T, Trampert P, Schneider L, Backes C, Ludwig N, Gerasch A, Kaufmann M, Gessler M, Graf N, et al. Multi-omics enrichment analysis using the GeneTrail2 web service. Bioinformatics. 2016;32(10):1502–8.
24. Keller A, Leidinger P, Vogel B, Backes C, ElSharawy A, Galata V, Mueller SC, Marquart S, Schrauder MG, Strick R, et al. miRNAs can be generally associated with human pathologies as exemplified for miR-144. BMC Med. 2014;12:224.

25. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, Wendschlag A, Giese N, Tjaden C, Ott K, et al. Toward the blood-borne miRNome of human diseases. Nat Methods. 2011;8(10):841–3.
26. Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, Berg D, Ruschil C, Galata V, Ruprecht K, et al. Validating Alzheimer's disease micro RNAs using next-generation sequencing. Alzheimers Dement. 2016:12(5): 565-76.
27. Backes C, Leidinger P, Altmann G, Wuerstle M, Meder B, Galata V, Mueller SC, Sickert D, Stahler C, Meese E, et al. Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. Anal Chem. 2015;87(17):8910–6.
28. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418–26.
29. Huang X, Yuan T, Tschannen M, Sun Z, Jacob H, Du M, Liang M, Dittmar RL, Liu Y, Liang M, et al. Characterization of human plasma-derived exosomal RNAs by deep sequencing. BMC Genomics. 2013;14:319.
30. Burgos KL, Javaherian A, Bomprezzi R, Ghaffari L, Rhodes S, Courtright A, Tembe W, Kim S, Metpally R, Van Keuren-Jensen K. Identification of extracellular miRNA in human cerebrospinal fluid by next-generation sequencing. RNA. 2013;19(5):712–22.

3.14 *Small ncRNA-Seq results of human tissues: variations depending on sample integrity*

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

This article is available under: https://doi.org/10.1373/clinchem.2017.285767

3.15 *Next generation sequencing analysis of total small noncoding RNAs from low input RNA from dried blood sampling*

This article is available under: https://doi.org/10.1021/acs.analchem.8b03557

This article is available under: https://doi.org/10.1021/acs.analchem.8b03557

This article is available under: https://doi.org/10.1021/acs.analchem.8b03557

This article is available under: https://doi.org/10.1021/acs.analchem.8b03557

This article is available under: https://doi.org/10.1021/acs.analchem.8b03557

This article is available under: https://doi.org/10.1021/acs.analchem.8b03557

3.16 *Comparative analysis of biochemical biases by ligation-and template-switch-based small RNA library preparation protocols*

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

This article is available under: https://doi.org/10.1373/clinchem.2019.305045

Published online 8 December 2020 *Nucleic Acids Research, 2021, Vol. 49, No. 2* **e10**
doi: 10.1093/nar/gkaa1122

CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing

Yongping Li[1,2,†], Tobias Fehlmann[1,†], Adam Borcherding[3], Snezana Drmanac[3], Sophie Liu[3], Laura Groeger[4], Chongjun Xu[2,3,5,6], Matthew Callow[3], Christian Villarosa[3], Alexander Jorjorian[3], Fabian Kern[1], Nadja Grammes[1], Eckart Meese[4], Hui Jiang[2], Radoje Drmanac[2,3,5,6], Nicole Ludwig[4,†] and Andreas Keller[1,7,*,†]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]MGI, BGI-Shenzhen, Shenzhen 518083, China, [3]Complete Genomics Incorporated, San Jose, CA 95134, USA, [4]Department of Human Genetics, Saarland University, 66421 Homburg, Germany, [5]BGI-Shenzhen, Shenzhen 518083, China, [6]China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China and [7]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA 94304, USA

Received May 15, 2020; Revised October 02, 2020; Editorial Decision November 02, 2020; Accepted November 26, 2020

ABSTRACT

Results of massive parallel sequencing-by-synthesis vary depending on the sequencing approach. CoolMPS™ is a new sequencing chemistry that incorporates bases by labeled antibodies. To evaluate the performance, we sequenced 240 human non-coding RNA samples (dementia patients and controls) with and without CoolMPS. The Q30 value as indicator of the per base sequencing quality increased from 91.8 to 94%. The higher quality was reached across the whole read length. Likewise, the percentage of reads mapping to the human genome increased from 84.9 to 86.2%. For both technologies, we computed similar distributions between different RNA classes (miRNA, piRNA, tRNA, snoRNA and yRNA) and within the classes. While standard sequencing-by-synthesis allowed to recover more annotated miRNAs, CoolMPS yielded more novel miRNAs. The correlation between the two methods was 0.97. Evaluating the diagnostic performance, we observed lower minimal *P*-values for CoolMPS (adjusted *P*-value of 0.0006 versus 0.0004) and larger effect sizes (Cohen's d of 0.878 versus 0.9). Validating 19 miRNAs resulted in a correlation of 0.852 between CoolMPS and reverse transcriptase-quantitative polymerase chain reaction. Comparison to data generated with Illumina technology confirmed a known shift in the overall RNA composition. With CoolMPS we evaluated a novel sequencing-by-synthesis technology showing high performance for the analysis of non-coding RNAs.

INTRODUCTION

Since the mid 1990′s, massively parallel sequencing approaches have been developed and continuously improved. The first commercial instruments were available on the market around 2005 (1). The rapid development of technology in the first 10 years had a substantial impact on genomic research (2), also leading to a continuous growth of data deposited in resources such as GenBank (3). While one of the most common applications is genome sequencing, RNAs are often analyzed using high-throughput sequencing as well. Even resolution at the single cell level can be reached now (4). A general overview of the different sequencing approaches together with available instruments highlights the diversity of available platforms and applications (5). Most recently, a comparison of Illumina NextSeq 500, NovaSeq 6000 and the BGI MGISEQ-2000 using identical single Cell 3′ libraries generated with the 10× Genomics Chromium platform highlighted comparable performance between the platforms in general (6).

For the high-throughput analyses of small non-coding RNAs (sncRNAs), sequencing has become one of the most frequently used methods (7). This has led to a very deep understanding of the sncRNA expression in humans (8,9) and many other species (10). As a consequence, databases on sncRNAs, especially on microRNAs (miRNAs) are updated regularly with increasing numbers of miRNAs. The miRBase in its most recent release 22 (October 2018 (11)) contains 38 589 entries from 271 species (12). Besides miRBase, MirGeneDB contains 10 899 curated miRNAs from

[*]To whom correspondence should be addressed. Tel: +49 681 302 68611; Fax: +49 681 302 68611; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first two and last two authors should be regarded as Joint First Authors.

45 different organisms (13) and miRCarta (14) has the ambition to provide a collection of all expressed small RNAs. With 11 000 annual publications on miRNAs, these databases cover particular needs of researchers and provide an important source of information for future miRNA annotations (15). The largest fraction of miRNAs from high-throughput sequencing has been annotated for *Homo sapiens*. For example, as of August 2020, the miRMaster web service (16) has been applied in over 1300 studies. Sequencing data of more than 74 000 human sncRNA samples were evaluated and 1.1 trillion reads (1.1×10^{12}) have been processed using miRMaster. Notably, only a fraction of all available sncRNA sequencing data has been analyzed using the miRMaster tool, e.g. since only one organism is considered. Thus, the total number of sncRNA sequencing data sets exceeds the figures given above substantially. The gold standard sncRNA analysis software miRDeep/miRDeep2 (17,18) for example has been cited almost 2000 times. Constantly decreasing cost and broader availability of sequencing systems will lead to a continuously growing amount of sncRNA datasets in the future.

Many studies, however, indicate a severe influence of sample handling, library preparation and the sequencing technology on the read quantity, composition and quality (19–22). The most commonly applied approach is sequencing-by-synthesis using Illumina systems. These are available in combination with different library preparation approaches (19). We previously evaluated the performance of sequencing-by-synthesis on Illumina systems to combinatorial probe-anchor synthesis (cPAS)-based BGISEQ-500 sequencer (23). As compared to the Illumina system we found a larger variety of sncRNAs in the cPAS data, including twice as much yet unknown microRNAs at that time. Both sequencing approaches however rely on similar sequencing-by-synthesis principles, incorporating labeled nucleotides during each sequencing cycle.

The continuous development of library preparation and sequencing approaches is leading to novel commercially available systems and assay formats. The availability of a new experimental approach however immediately calls questions with respect to the validity of its data and the comparability. Especially for applications in biomarker development a platform change may significantly affect the diagnostic or prognostic performance of tests. Consequently, two questions come up whenever a new experimental approach is available: how does the performance change if technical replicates are compared between platforms and how does it affect biological results?

Recently, a fundamentally novel sequencing approach called CoolMPS has been introduced and made commercially available through MGI Tech Co., Ltd, Shenzhen, China (details are provided in the 'Materials and Methods' section). While it still relies on the sequencing-by-synthesis principle as other methods, no labeled nucleotides are incorporated. In order to measure a signal intensity representative for the incorporated base at each cycle, four specific antibodies, one recognizing each of the four natural bases (A, T, C, G) are used. The approach promises higher data quality by avoiding incorporation and detection interference of base-linked dyes and providing stronger signals by attaching multiple molecules of a dye per antibody. The CoolMPS approach for sequencing non-coding RNAs is described in the 'Materials and Methods' section. More details on the sequencing kits and basic biochemical principles of the methodology and its application are available with the user manual of the commercial kits and as preprint (https://doi.org/10.1101/2020.02.19.953307). It is mandatory to evaluate such new technologies with respect to common application scenarios. Discovering single nucleotide variants or small insertions and deletions pose different challenges as compared to, e.g. the quantification of RNAs in an at least pseudo-quantitative manner. In this study, we set to present the first detailed and direct performance comparison between the novel antibody-based labeling approach in comparison to standard sequencing-by-synthesis using labeled nucleotides for the quantification of small non-coding RNAs.

MATERIALS AND METHODS

RNA sample preparation and quality control

RNA from 2.5 ml whole blood collected in PAXgene tubes was isolated using the PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany) according to the manufacturer's recommendations. RNA concentration and integrity were measured using Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and RNA 6000 Nano Kit for Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), respectively. RNA was aliquoted and used for the four experimental approaches CoolMPS, BGISEQ, Illumina and reverse transcriptase-quantitative polymerase chain reaction (RT-qPCR) as described below. The study was approved by the ethical committee of the Medical Faculty of the University of Tuebingen (Nr. 90/2009BO2). A list of samples included in the study is available as Supplementary Table S1.

CoolMPS™ on the DNBSEQ-G400RS

MiRNA libraries were prepared using the MGIEasy Small RNA Library Prep Kit (MGI Technologies, Shenzhen, China; product number 1000006383) with 800 ng total RNA input according to the manufacturer's recommendations. First, adapter sequences were ligated to the 3′ end of the RNA, followed by ligation of barcoded RT primers. Next, a universal adapter was ligated to the 5′ end. The RNA was then transcribed into cDNA by HiScript II Reverse Transcriptase in the presence of RNAse inhibitor. The primers used for the reverse transcription contained barcodes that allowed the pooling of up to 24 samples per sequencing library. Then cDNA libraries were amplified by 18-cycles of PCR reactions. Amplified PCR products were size selected using 6% TBE gel electrophoresis and the band from 100 to 120 bp was then purified with spin-X centrifuge tube filters followed by ethanol precipitation. The purified PCR products were quantified using Qubit dsDNA HS Assay kit (Invitrogen, Cat No. Q32854). Twelve purified PCR products were pooled with 84 fmol each (total 1 pmol) and circularized using a specific oligo sequence complementary to sequences in both the 3′ and 5′ adaptors provided in the MGIEasy Small RNA Library Prep Kit. The remaining linear DNA was digested.

After purification, the single strand circularized DNA library was quantified using Qubit ssDNA Assay Kit (Invitrogen, Cat No, Q10212). Subsequently, DNA nanoballs (DNBs) were generated using rolling circle amplification from 60 fmol of single stranded, circularized DNA library for 25 min. The DNB concentration was determined using Qubit ssDNA Assay Kit. The DNBs (concentration in the range of 8–20 ng/μl) were mixed with loading buffer by manual pipetting and subsequently loaded onto DNBSEQ-G400RS 4-lane flowcells (product number 1000016985) using the MGIDL-200H DNB loader as described in the CoolMPSTM High-throughput Sequencing Set User Manual provided with the kit. Loaded flow cells were sequenced on the DNBSEQ-G400RS instrument using CoolMPS^{TM} SE50 beta sequencing kits, now available as commercial products (product number 1000019478, MGI Tech Co., Ltd, Shenzhen, China) following manufacturers recommendation. The MGI CoolMPS^{TM} SE50 kits are the standard product for small RNA sequencing. Sequencing was performed by selecting the smallRNA sequencing plan from the application menu on the DNBSEQ-G400RS. Single end sequencing of 50 bp along with 10 bp of barcode was performed. The basic difference between CoolMPS and standard sequencing-by-synthesis, relying on incorporation of labeled nucleotides, is the incorporation of unlabeled, reversibly terminated nucleotides. The fluorescent signal to detect the incorporated bases is generated by using base-specific 3′ block-dependent fluorescently labeled antibodies. After each cycle, the bound antibodies are removed and 3′ blocking moiety on the sugar group of the nucleotide regenerates the natural nucleotides. This procedure has the advantage not leaving a mark on the base and making the current sequencing cycle independent on the previous one. Base calling and generation of FASTQ files on the DNBSEQ-G400RS was performed using the software release for CoolMPS (BasecallLite version_1.0.7.84). An important machine quality control step included the removal of tiles from the FASTQ files that failed at some point in the base calling process leading to 'N' bases for all reads in that respective tile. A detailed description of the CoolMPS method and procedures is available under: https://doi.org/10.1101/2020.02.19.953307. The sequencing has been performed by Complete Genomics Incorporated, San Jose, California. The overall process of library preparation and sequencing on the DNBSEQ-G400 is referred to as 'CoolMPS' through the whole manuscript.

BGISEQ-500 sequencing using standard cPAS

As described above for CoolMPS, the MGIEasy Small RNA Library Prep Kit (product number 1000006383) was used to generate circularized DNA libraries with 800 ng total RNA input according to the manufacturer's recommendations. The library preparation and DNB preparation procedures are exactly the same as the one described in the previous section. DNBs were loaded onto the flow cell using the BGIDL-50 DNB loader and single end 50 bp sequencing was performed using the BGISEQ-500RS High-throughput Sequencing Set SE50 on the BGISEQ-500RS instrument. The sequencing has been carried out in the Human Genetics Department at Saarland University, Germany. This process is referred to as 'BGISEQ' through the whole manuscript.

Illumina library preparation and sequencing

Libraries were prepared according to the protocol of the TruSeq Small RNA Sample Prep Kit (Illumina) with 200 ng of total RNA per sample as starting material as described previously (24). In brief, the concentration of the libraries was assessed using a Bioanalyzer with the DNA 1000 Chip. Before sequencing, libraries were pooled in equal amounts of batches of six samples and clustered with a concentration of 9 pmol in one lane each of a single read flow cell. Sequencing of 50 cycles was performed on a HiSeq instrument (Illumina). Demultiplexing of raw sequencing data and generation of FASTQ files was performed with CASAVA v1.8.2.

RT-qPCR

RT-qPCR experiments are described in detail in the original publication (25). In brief, the miScript PCR system was used with custom miRNA PCR arrays (all reagents from Qiagen, Hilden, Germany). The PCR arrays were designed in 96-well plates to measure the expression of human miRNAs and RNU48 as well as RNU6 as endogenous controls. The RT-qPCR experiments have been performed in the Human Genetics Lab of Saarland University. Reverse transcription was performed using 100 ng total RNA as input using miScriptRT-II kit in 20 μl total volume. PCR reactions with 1 ng cDNA input in a total volume of 20 μl were set up automatically using the miScript SYBR Green PCR system in a Qiagility pipetting robot (Qiagen, Hilden, Germany).

Bioinformatics

The pre-processing of the FASTQ files of CoolMPS, BGISEQ and Illumina has been done using miRMaster 1.1 (16,26). MiRMaster is freely accessible at https://www.ccb.uni-saarland.de/mirmaster/. Briefly, adapters at the 3′ end were trimmed, while allowing an error of maximum one base and requiring a minimum overlap with the read of 10 bases. Reads were quality trimmed when the average quality dropped below 20 in a window of four consecutive bases to ensure a high quality of reads used for the downstream processing. All reads shorter than 17 bases after trimming were discarded from all further analyses. Read duplication levels were computed with FASTQC 0.11.8. The error rate per base was estimated by mapping the trimmed reads to the human genome with bowtie, while allowing up to three mismatches (command line: bowtie -v3 -k 1 –best –fullref) and counting the mismatched bases with Samtools stats (version 1.9, (27)). To further ensure the best comparability, BGISEQ and Illumina data were subsampled to match the CoolMPS distribution that was originally sequenced to a lower extent. In detail, all samples were subsampled to a read depth of 10 Million reads. Reads were mapped to the primary assembly of GRCh38.p10 using bowtie 1.2.2 (28), while allowing no mismatches and discarding reads mapping to over 100 locations (command line: bowtie -v0 -m 100 –best –strata –fullref). Read RNA classes were determined using FeatureCounts 1.5.2 (29) and annotations of

GENCODE v25 (30), piRBase 1 (31), miRBase v22.1 and GtRNAdb 18.1 (32) with the following parameters: -F SAF –O –M –R –f –fracOverlap 0.9, which required an overlap of at least 90% of a read with the annotated region and allowed multimapping reads and overlapping features. MiRBase v22.1 miRNAs were quantified using miRMaster with up to one mismatch and a variability of two bases allowed at the 5′ end and five bases at the 3′ end. Novel miRNA candidates were predicted with miRMaster with a required minimum expression of five reads in at least 75% of all dementia or control samples. Since we expect numerous false positive hits from the next generation sequencing data we performed a quality control of the newly predicted candidates and evaluated them using the NovoMiRank tool (33). NovoMiRank was applied using the default parameters, i.e. miRBase versions 1–7 were used as reference to identify the most reliable candidates. All further downstream analyses have been carried out in R 3.6.1 (https://www.R-project.org/). To test whether miRNAs were normally distributed, Shapiro–Wilk tests were computed per miRNA using the shapiro.test function from the stats package. As hypothesis test, parametric *t*-test and non-parametric Wilcoxon Mann-Whitney (WMW) test were performed using the t.test and wilcox.test functions from the stats package. Statistical tests for group comparisons were carried out as two-tailed and un-paired tests. All *P*-values were subjected to adjustment for multiple testing by using the Benjamini–Hochberg approach through applying the p.adjust function from the stats package. To estimate the effect sizes, the area under the receiver characteristic curve (AUC value) and the Cohen's D effect size were computed using the R pROC package (1.15.0, (34)) and the R effsize package (0.7.4). Plots were generated with ggplot2 (3.1.0), cowplot (0.9.4), complexHeatmap (2.5.3, (35)), ggridges (0.5.1) and vioplot (0.3.5). To compute the most significant overlap between the CoolMPS and BGISEQ technology in terms of dementia biomarkers we employed the dynamic programming based DynaVenn approach (36). DynaVenn is freely accessible at https://www.ccb.uni-saarland.de/dynavenn. Functional categories were analyzed by miRNA set enrichment analysis with default parameters using miEAA 2.0 (37,38) with a list of the miRNAs sorted with respect to their effect sizes as input (with separate adjustment of categories and Benjamini–Hochberg adjustment procedure).

RESULTS

Study setup allowing to evaluate technical and biological aspects

Primary aim of the study was to compare the combinatorial probe-anchor synthesis (cPAS)-based data using labeling of nucleotides to the data generated by the new antibody labeled-based approach on the more recent DNBSEQ-400RS systems. In the context of this manuscript, the former approach is referred to as BGISEQ and the latter as CoolMPS. Secondary aim was to compare the performance and comparability of both approaches in terms of potential liquid biopsy biomarker tests. We thus selected a study setup that allows to address both aims (Figure 1A). We sequenced 240 individual blood samples on both sequencing systems. The 240 samples include 179 controls and 38 patients with dementia. This part of the cohort has been used to evaluate the performance of both technologies to detect dementia biomarkers. Furthermore, the 240 samples include 17 individuals and 6 technical replicates. The latter samples were not used for the biomarker study but to assess the general stability and reproducibility of the technologies. Further, we compared the data to RT-qPCR measurements of a subset of 19 miRNAs in 189 samples and also evaluated the performance in comparison to data generated by Illumina sequencers. A full list of miRNAs and samples together with the respective Delta CT values from the RT-qPCR validation is available in Supplementary Table S2. We first evaluated the general performance of CoolMPS for quantification of RNA and then provide results of CoolMPS as liquid biopsy biomarker for dementia. The cohort was composed of participants with an average age of 67.3 years and a standard deviation of 12.3 years (Figure 1B). Details on the sequencing approaches and data analyses are given in the 'Materials and Methods' section.

Key performance indicators reveal improved quality of CoolMPS

First, we compared the Q30 values for the reads from the two sequencing approaches (Figure 1C). The Q30 value provides the percentage of bases sequenced with a Phred score of at least 30, corresponding to an error rate of 0.1%. The median Q30 of the BGISEQ was 91.8% while the median Q30 of CoolMPS jumped to 94%, representing a significant improved performance of CoolMPS ($P < 10^{-10}$). Intriguingly, we observed the higher per base sequencing accuracy over the complete read length not observing any drop at the beginning or at the end of the read. Moreover, CoolMPS showed lower variability in sequencing performance over the read in general as well as lower variability per base in the read (Figure 1D). While the variation of valid reads per sequencing run still varied for the CoolMPS technology we observed a constantly higher fraction of reads mapping without mismatches to the human genome (84.9% for BGISEQ and 86% for CoolMPS; Figure 1E). We also investigated the GC content of the generated libraries and found a median of 51.10% for BGISEQ and a median of 50.72% for CoolMPS in the unprocessed data, which dropped to a median of 42.38 and 41.60% for BGISEQ and CoolMPS after adapter and quality trimming, respectively (Supplementary Figure S1A and B). The mean quality scores per position varied between 33.95 and 36.35 for CoolMPS and even increased slightly toward the end of the read. In contrast, the BGISEQ reads varied between 27.95 and 36.17 and reached their peak at position 26. Then, the quality of BGISEQ reads decreased until position 50 (Supplementary Figure S1C and D). The mean quality scores for the trimmed files, i.e. those that did not contain any adapters, varied similarly, although the mean quality scores decreased more for longer reads. The estimated error rate was for both technologies similar with a median of 0.74% for BGISEQ and 0.76% for CoolMPS (Supplementary Figure S1E). For both, the raw sequencing files, and the trimmed ones, we observed a close to identical GC content distribution. For both technologies we observed two distinct peaks at 51 and 57% (Supplementary Figure S1F and G). We also found that the

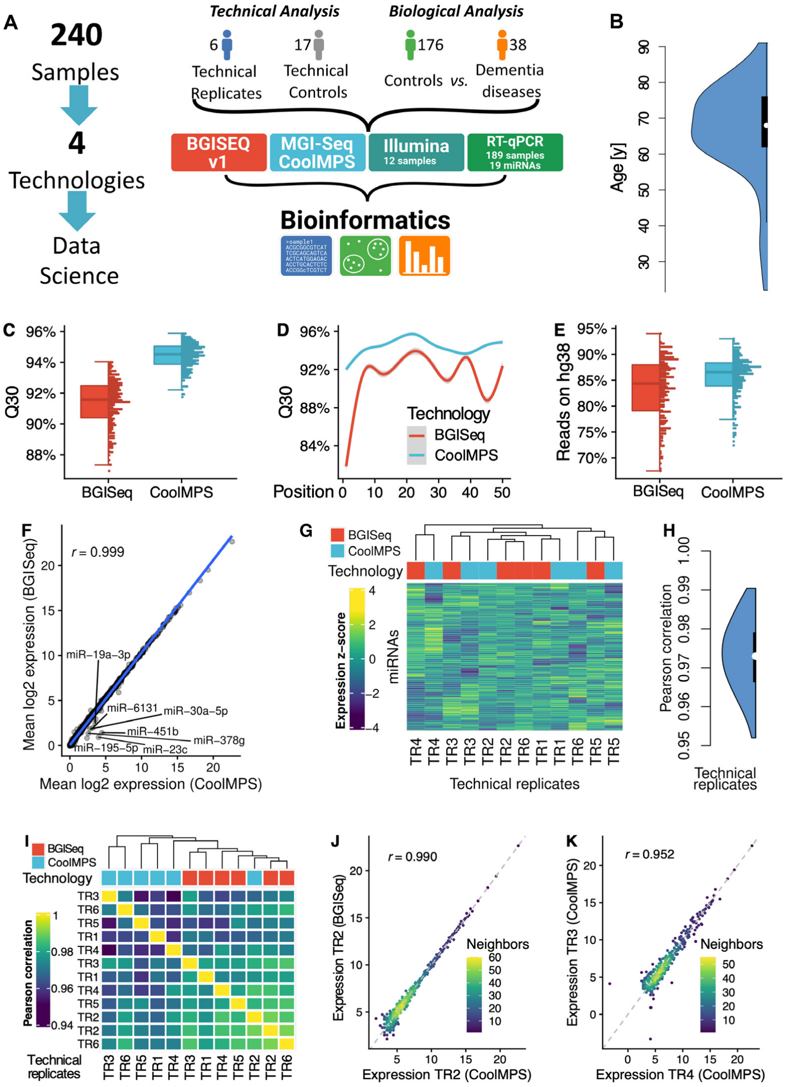


Figure 1. Study setup and quality control. (**A**) In the study we measured 240 individual blood samples using two fundamentally different sequencing approaches and compare the data by bioinformatics approaches before we compute the concordance to RT-qPCR profiles. The 240 samples include one part that has been used only for assessment of technical properties (6 and 17 samples in blue and gray) as well as a second part to evaluate performance related to biomarker discovery (176 controls in green and 38 dementia cases in orange). (**B**) Distribution of the age of the individuals included in the study, shown as violin plot. The black box spans the first to the third quartile and the white dot shows the median. (**C**) Distribution of the average Q30 value per sample for the two technologies, shown as boxplot (left) and dotplot (right). Each sample is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**D**) Q30 value over all samples per technology as function of the position in the read. The smoothed curve is fitted by a generalized additive model using a cubic regression spline. The gray area represents the confidence interval of the fit. (**E**) Distribution of the percentage of reads mapping to the human reference genome hg38 without mismatch per technology, shown as boxplot (left) and dotplot (right). Each sample is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**F**) Scatter plot of the average expression of all miRNAs in all samples for the two technologies. The blue line is the regression line. The Pearson correlation is shown in the upper left part of the plot. MiRNAs with a fold change larger than two between both technologies are highlighted. (**G**) Heat map of the clustered expression *z*-scores of miRNAs (rows) and technical replicates (columns). The color code for the columns represents the technology. The dendrogram shows the hierarchical clustering of the samples with Euclidean distance and complete linkage. (**H**) Distribution of all $12*11/2 = 66$ pairwise Pearson correlation coefficients, shown as violin plot. The black box spans the first to the third quartile and the white dot shows the median. (**I**) Correlation matrix of the expression values of all miRNAs for all technical replicates. The dendrogram shows the hierarchical clustering of the samples with Euclidean distance and complete linkage. (**J**) Scatter plot of miRNAs for the best correlation between two technical replicates. The dotted line represents the angle bisector. The Pearson correlation is shown in the upper left part of the plot. The points are colored according to the point density in their neighborhood. (**K**) Scatter plot of miRNAs for the worst correlation between two technical replicates. The dotted line represents the angle bisector. The Pearson correlation is shown in the upper left part of the plot. The points are colored according to the point density in their neighborhood.

read length in both libraries after trimming peaked at 22, as we expected from a miRNA enriched library (Supplementary Figure S1H). We further evaluated the duplication levels of the CoolMPS and BGISEQ libraries. In both cases, the distributions were again nearly identical, showing most duplication levels above 10 000 (Supplementary Figure S1I and J). This is expected from miRNA libraries, as often a small number of miRNAs account for most of the reads. Finally, we checked the read base composition and found similar patterns. The first 22 bases reveal the most overrepresented sequence (i.e. the sequence of hsa-miR-451a), followed by the bases of the adapter sequence for the raw reads, and by less sequence specific bases for the trimmed reads (Supplementary Figure S1K and L). For most of the tested relevant key performance indicators (e.g. Q30 and reads mapping to the human genome) that allow to compare the general sequencing performance, CoolMPS yielded an increased performance compared to the classical BGISEQ approach.

Next, we evaluated and compared the reproducibility of the two technologies. When comparing the mean expression of all samples for CoolMPS to BGISEQ we obtained an extremely high correlation of 0.999 (Figure 1F). The scatter plot highlights a set of seven miRNAs, which were measured with higher expression in the CoolMPS data as compared to BGISEQ (miR-19a-3p, miR-30a-5p, miR-6131, miR-451b, miR-378g, miR-195-5p and miR-23c). Next, we considered only the six technical replicates per technology. There, these miRNAs reveal the same pattern as for the complete set of samples, thus excluding variance related to the disease status of the participants as potential cause (Supplementary Figure S2). Sequence and structure properties of these miRNAs are shown in Supplementary Table S3. Neither the length, nor the base composition or secondary structures reveal a joint pattern, arguing against a technological bias. We then asked whether we observe a clustering according to the sequencing approach or whether CoolMPS and BGISEQ samples mix. Indeed, hierarchical clustering indicates that the samples do not cluster by technology (Figure *1G*). The Pearson correlation between all $12 \times 11 / 2 = 66$ pair wise comparisons of technical replicates varied between 0.952 and 0.990 with a median performance of 0.973 (Figure 1H). The correlation matrix revealed marginal differences in the correlation coefficients between all the BGISEQ replicates (median 0.980) in comparison to the ones between the CoolMPS samples (median 0.964) (Figure 1I). Also, the correlation between the two technologies with a coefficient of 0.973 was high. The differences in the correlation lead to a tendency of technologies to cluster together, although CoolMPS Technical Replicate 2 clustered with BGISEQ Technical Replicates 2 and 6. Scatter plots for the best (Figure 1J) and the worst correlation (Figure 1K) demonstrate the generally very high reproducibility between the technologies that is in the same range as technical replicates within the technologies. Most importantly, we did not observe any significant change between the RNAs profiled with BGISEQ compared to CoolMPS after adjustment for multiple testing, both, for the WMW and the *t*-test.

Having understood basic performance of the sequencing technology as well as core aspects on technological reproducibility we next evaluated the content of the different sequencing approaches with respect to quantitative and qualitative aspects.

Composition of different RNA classes is similar between BGISEQ and CoolMPS

The first question related to small non-coding RNA sequencing data is the representation of different RNA classes. Different sample- and library preparation protocols lead to varying results. For example, size selection is applied to enrich-specific populations of sncRNAs. To minimize respective effects and to focus on the performance of the sequencing technique, we used the same libraries for sequencing and purified small non-coding RNAs by gel electrophoresis (see 'Materials and Methods' section). This protocol has been optimized to enrich for miRNAs, however, leaving also reads to evaluate other RNA classes. The distribution to the different classes matched generally very well between BGISEQ and CoolMPS (Figure 2A). Especially, we observed the intended enrichment for miRNAs. For BGISEQ, 91.7% of all mappable reads matched to miRNAs, for CoolMPS we even reached a higher mapping of 92.7%. The second most abundant RNA class was the Ensembl's misc RNA category, containing among others yRNAs and signal recognition particle RNAs (SRP RNAs). This category contains 5.1% of all BGISEQ and 4.5% of all CoolMPS reads. All other categories were covered by less than 1% of reads in both technologies. The scatter plot contrasting the $\log_{10}$ percentages for both technologies highlights the very reproducible distribution of reads to the different RNA classes (Figure 2B). Since the protocol was optimized to enrich for miRNAs and our results demonstrate that this enrichment was successful, we focused on comparing the performance for this class of sncRNAs.

CoolMPS yields more novel miRNA candidates

With respect to different technologies a bias in sncRNA-seq data is known. Especially for specimen types such as whole blood where already an enrichment of selected miRNAs exist, additional technological bias can further impair the data analysis. In whole blood, miRNA expression is not uniformly distributed but few miRNAs are significantly higher expressed than others. Technology bias further overamplifies the respective miRNA reads. These circumstances complicate the discovery of new miRNAs with the aim of completing the repertoire of annotated miRNAs (8). We thus evaluated and compared the distribution of reads to different miRNAs using the two sequencing technologies and asked how many novel miRNA candidates could be obtained. As expected, we observed an uneven distribution, which is however highly concordant between the technologies (Figure 3A). At the same time, we discovered 124 novel miRNA candidates using BGISEQ while CoolMPS based results highlight 134 novel miRNA candidates (Figure 3B and Supplementary Figure S3A). These findings suggest a higher sensitivity in terms of discovering low abundant yet unknown miRNA molecules. Remarkably, a large fraction of all new microRNA candidates, in total 88, have been detected by both technologies. To assess the quality of those

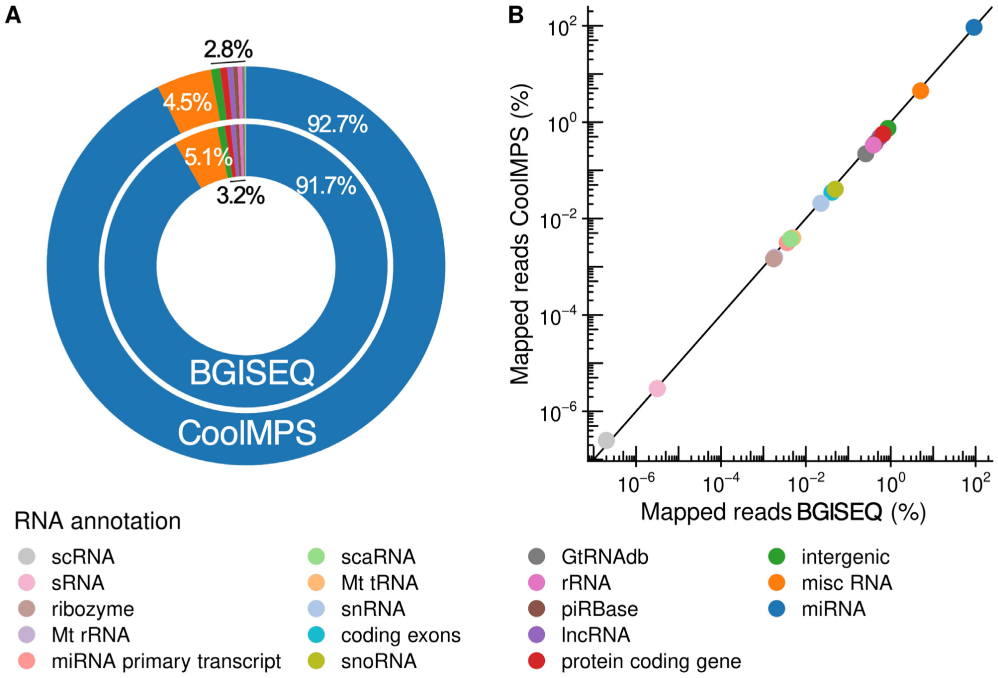


Figure 2. Distribution to the different sncRNAs classes. (**A**) Donut plot comparing the distribution of all RNA classes and intergenic regions that were covered by reads from CoolMPS and BGISEQ. (**B**) Scatter plot that shows the percentage of reads mapping to the RNA classes and intergenic regions for BGISEQ (*x*-axis) and CoolMPS (*y*-axis) on a logarithmic scale.

miRNA candidates we scored them using NovoMiRank. The score computed by NovoMiRank considers sequence and structural features and describes the average distance of the new candidates to a reference set, which is per default miRBase v1-7. The median score obtained for the common candidates was 1.12, while the technology specific candidates obtained median scores of 1.05 for CoolMPS and 1.00 for BGISEQ. As highlighted by the distribution shown in Supplementary Figure S3B, the score ranges are similar between the approaches and only few candidates (four detected by both CoolMPS and BGISEQ, three CoolMPS specific and one for BGISEQ specific) showed scores above 1.5. The score of 1.5 has been set since it is the maximum score observed for miRBase v1-7 miRNAs i.e. the reference set of NovoMiRank. In summary, both technologies do not reveal quantitative differences in the quality of reported miRNAs but only in the quantity, with remarkable advantages of CoolMPS.

In comparing the distribution of miRNAs annotated in the miRBase we observe 76.7% of all BGISEQ reads mapping to the most abundant miRNA (miR-451a; Figure 3C). Using CoolMPS, 78.5% of all reads matched to this miRNA (Figure 3D). The second most abundant miRNA is represented by 9 and 8.4% of all reads, respectively (miR-92a-3p). In sum, the top five miRNAs are covered by 93.8% of all reads in the BGISEQ and by 92.6% of all reads in the CoolMPS approach. A more detailed breakdown by excluding the most abundant miR-451a demonstrates that the order of the 10 most abundant miRNAs matches perfectly between the two technologies (Figure 3E and F). At the same time, the data reinforces that especially for biospecimens with an uneven distribution of miRNA molecules, deep sequencing with the least possible bias is required to profile known and to discover new miRNAs.

Comparing biomarker profiles shows high reproducibility between the different approaches

One of the most important question in introducing new technologies is not only whether general performance improves but also whether previous biological results can be reproduced. One core example are biomarker tests. Often, biomarker sets change substantially when a new quantification approach is introduced. This might be an expected and even desired result, e.g. if a new technology generation with higher technical sensitivity is introduced. But if a new technology has the main task to support translation of biomarkers to care by facilitating better integration into clinical workflows or lower experimental costs, original biomarker profiles should not be compromised. We thus evaluated the diagnostic performance of miRNA biomarkers using BGISEQ and CoolMPS and used a liquid biopsy dementia test as validation example. We sequenced cases with dementia as well as controls with similar age distribution (Figure 1A and B). As performance criteria we considered the result of two commonly used hypothesis tests, the *t*-test and the WMW test. Since not all miRNAs were normally distributed according to the Shapiro Wilk test, we here focus on the results of the WMW test and provide the t-test *P*-values only in the supplement (Supplemen-

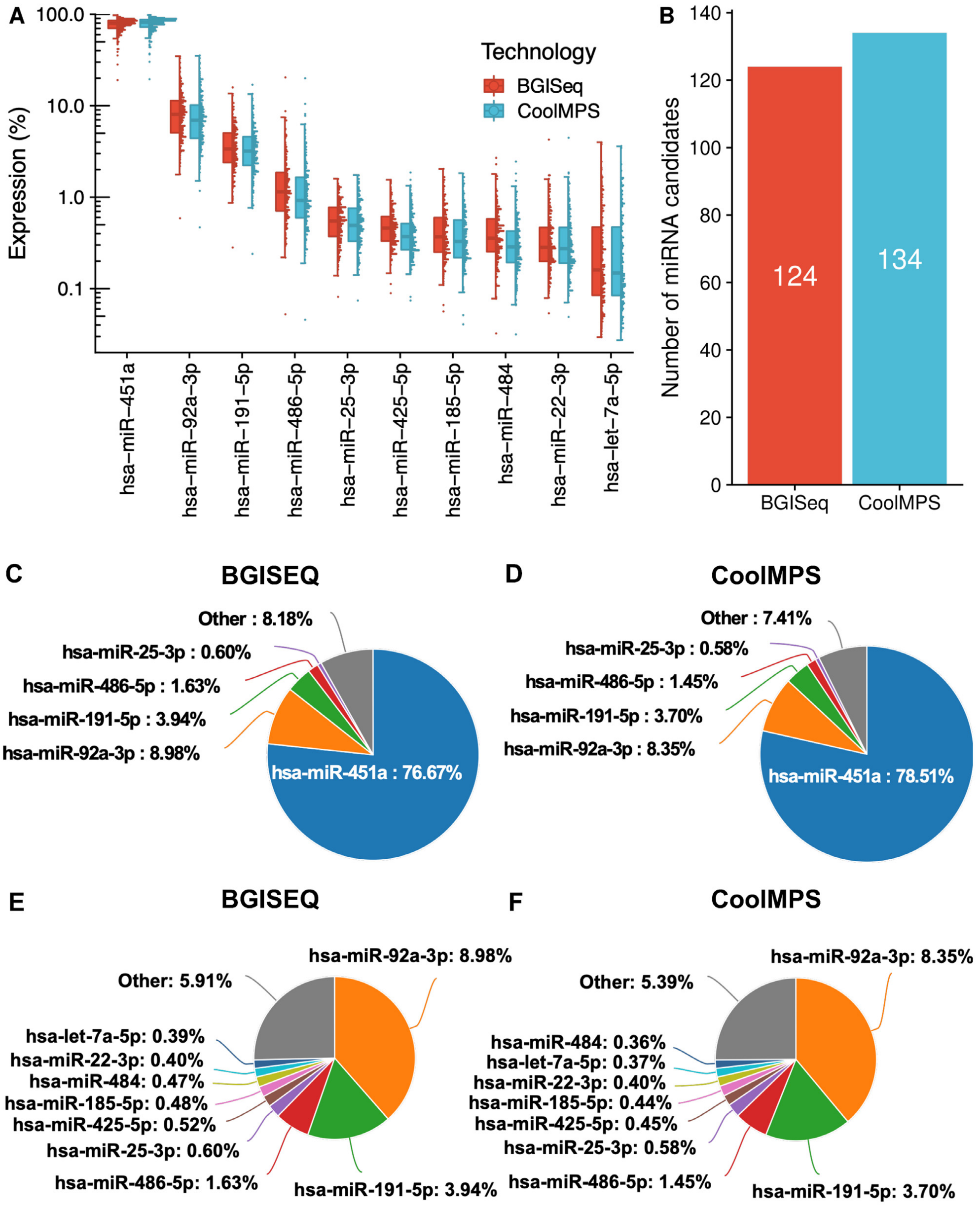


Figure 3. Distribution to microRNAs. (**A**) Distribution of the read percentage of the 10 most abundant miRNAs in the CoolMPS and BGISEQ data, shown as boxplot (left) and dotplot (right). Each sample is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**B**) Number of novel microRNA candidates for both technologies. (**C**) Pie chart for the top five miRNAs on the BGISEQ. (**D**) Pie chart for the top five miRNAs on the CoolMPS. (**E**) Pie chart for the top ten miRNAs on the BGISEQ after exclusion of the most abundant miR-451a. (**F**) Pie chart for the top ten miRNAs using CoolMPS after exclusion of the most abundant miR-451a.

tary Table S4 and 5). Because of known challenges with *P*-values and the controversial discussion on this topic (39), we also computed effect sizes, namely Cohen's D and the area under the receiver characteristics curve AUC. Detailed results for each miRNA and each of the different metrics are provided for both BGISEQ (Supplementary Table S4) and CoolMPS (Supplementary Table S5). In terms of AUC, BGISEQ and CoolMPS showed an almost identical distribution (Figure 4A). The scatter plot displays a very high degree of reproducibility (Pearson correlation coefficient of 0.905) between the two technologies considering the diagnostic performance (Figure 4B). As consequence, also the volcano plots for the two technologies were very similar (Figure 4C and D). Given the general concordance of

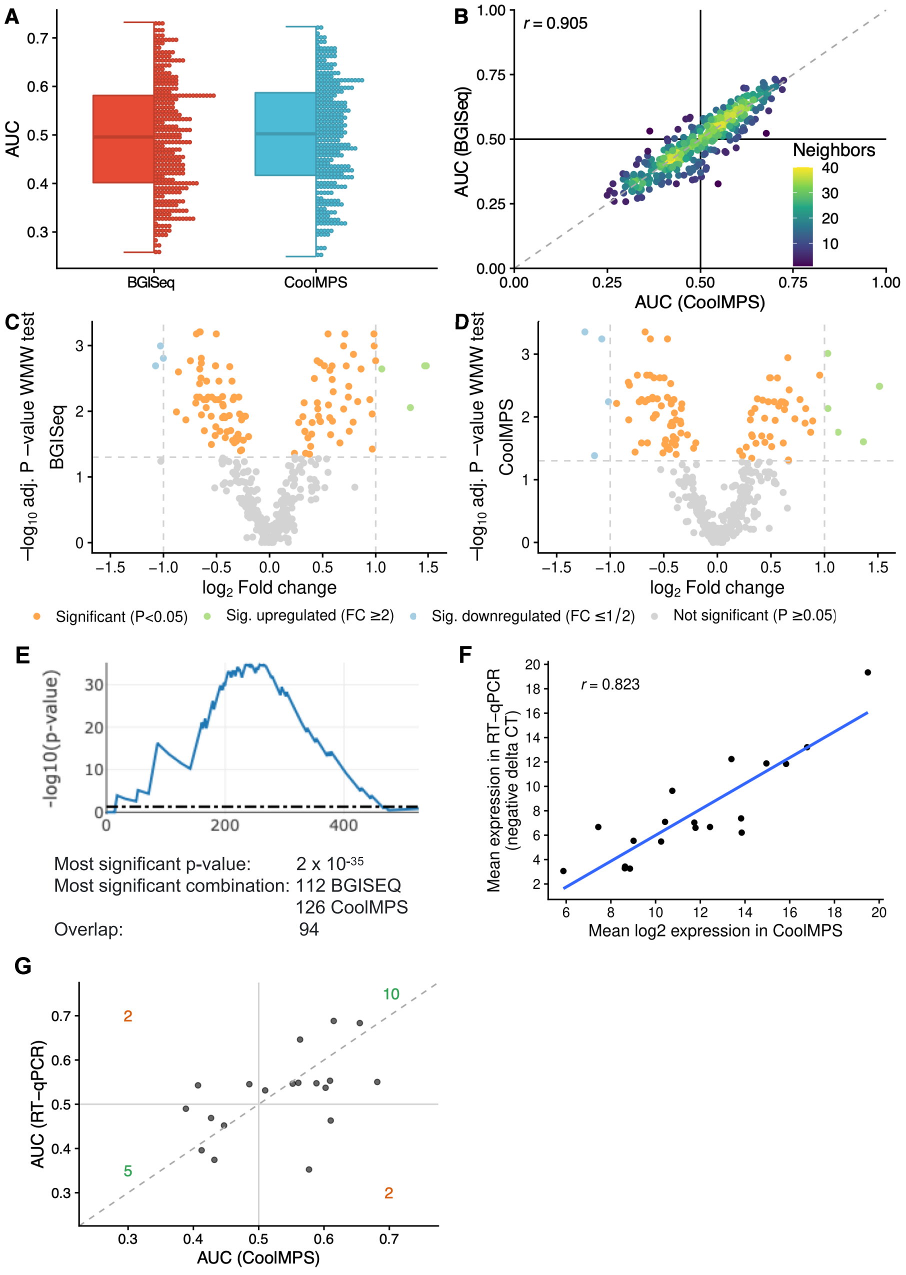


Figure 4. Diagnostic performance on dementia patients. (**A**) Distribution of the AUC values to differentiate between dementia and controls obtained for both technologies. An AUC of 0.5 means no dys-regulation. A deviation from 0.5 toward one means an upregulation and toward zero a downregulation of the biomarkers. The distribution is shown as boxplot (left) and dotplot (right). Each miRNA is shown as one dot. The boxes span the first to the third quartile with the horizontal line inside the box representing the median value. The whiskers show the minimum and maximum values or values up to 1.5 times the interquartile range below or above the first or third quartile if outliers are present. (**B**) Scatter plot of the AUC values to differentiate between dementia and controls in CoolMPS (*x*-axis) versus BGISEQ (*y*-axis). The black horizontal and vertical line represent the AUC value of 0.5, respectively. The Pearson correlation is shown in the upper left part of the plot. The points are colored according to the point density in their neighborhood. (**C**) Volcano plot showing the $\log_2$ fold change on the x-axis and the FDR adjusted negative $\log_{10}$ of the Wilcoxon–Mann–Whitney (WMW) *P*-value on the *y*-axis for BGISEQ. Orange dots are located above the horizontal line and are significant. Blue and green dots above the horizontal and on the left / right of the vertical lines are significant and have a fold-change above 2. (**D**) Same volcano plot as in Figure 4C, but for CoolMPS. (**E**) Result of DynaVenn that presents the negative $\log_{10}$ of the overlap between the two miRNA sets dependent on how many miRNAs are included. The peak of the curve represents the most significant overlap. (**F**) Scatter plot of the $\log_2$ CoolMPS expression (*x*-axis) and the negative delta CT value for the 19 miRNAs included in the validation study. The Pearson correlation coefficient is shown in the upper left part of the plot. (**G**) Scatter plot of the AUC values to differentiate between dementia and controls in CoolMPS (*x*-axis) and BGISEQ (*y*-axis). The dashed line represents the angle bisector.

the results we speculated that also the ranks of biomarkers were consistent between the two technologies. For the top-10 markers of BGISEQ and CoolMPS we thus compared the ranks and absolute values (Table 1). First, we recognized that the top marker performed better in CoolMPS as compared to BGISEQ in all metrics, the raw *P*-value, the adjusted *P*-value, the Cohen's D and the AUC. The adjusted *P*-values were for example 0.0006 in BGISEQ data and 0.0004 in CoolMPS data. Second, we observed that four miRNAs were among the top 10 markers in both technologies (miR-3200-3p, let-7e-5p, miR-15b-5p, miR-19b-3p). For other markers we computed partially very different ranks. One of the most extreme examples is miR-3335-5p, which is ranked 9th most significant in CoolMPS and 117th in BGISEQ. Nonetheless, this miRNA was significant in both approaches. On the one hand we observed a very high correlation, on the other hand, we also noticed substantial differences in the ranks, most likely related to the close range of the *P*-values, challenging the concept of fixed thresholds. To overcome the bias of selecting fixed rank ranges, we developed the DynaVenn approach that computes the most significant overlap between two biomarker sets containing technical or biological replicates. DynaVenn computed the best overlap in selecting the best 112 miRNAs from BGISEQ and the best 126 miRNAs from CoolMPS, yielding an overlap of 94 miRNAs and a *P*-value of 2×10^{-35} (Figure 4E). Thus, the two biomarker sets show a highly significant overlap which might have remained hidden if only the top 10 markers would have been considered.

Illumina sequencing data shows differently biased but comparable measurements

In addition to BGISEQ we also compared the performance of CoolMPS to standard Illumina sequencing for small non-coding RNAs on a subset of 12 samples (24). As part of our quality control we filter reads shorter than 17 nucleotides. We thus compared the fraction of filtered reads for the three technologies on the subset of samples sequenced by the three technologies. For BGISEQ, 2.76% (SD of 0.56) of reads, for CoolMPS 3.90% (SD of 0.54) of reads and for Illumina 3.95% (SD of 3.24%) of reads were excluded. In a first analysis step we evaluated the Q30 values obtained by both approaches and found a median Q30 of 94.95% for Illumina in comparison to 93.10% for CoolMPS (Supplementary Figure S4A). The quality profile revealed Q30 values going up to a median of 99.43% for Illumina in the first 20 positions, whereas a strong drop could be observed afterwards, going down to a Q30 of 87.33% at position 50 (Supplementary Figure S4B). In comparison, the CoolMPS quality remained more stable for the complete read length with an average Q30 of 93.12% (SD: 1.79%) and even showed an increased quality toward the end of the reads. For the fraction of reads that can be used in further analyses, i.e. the ones mapping to the human genome, we observed for CoolMPS a median of 90.74%, while for Illumina only 77.85% could be mapped (Supplementary Figure S4C). In the next step, we inspected the expression similarity of both technologies and found a general agreement of both with a Pearson correlation of 0.873 (Supplementary Figure S4D). Nevertheless, we could observe 52 miRNAs with expression values differing by fold changes above 10, showing the technological specific biases (e.g. hsa-miR-486-5p was expressed 76 times higher in the Illumina samples). In addition, we confirmed that the samples of both technologies clustered separately according to their miRNA profiles (Supplementary Figure S4E) and showed a much higher intra-technology expression correlation (Pearson correlation of 0.960 for CoolMPS on median, 0.955 for Illumina) than between technologies (median Pearson correlation of 0.742) (Supplementary Figure S4F and G). We then asked if the RNA class distribution between both technologies show similar patterns. We found that the CoolMPS samples showed a higher diversity of RNA classes, whereas the Illumina samples contained a higher percentage of reads mapping to piRNAs (0.70 versus 0.17% in CoolMPS) and miRNAs (97.76 versus 94.98% in CoolMPS) (Supplementary Figure S5). Next, we focused on the composition of the detected miRNAs and found that of the top 10 most expressed miRNAs of both technologies, six overlapped. The largest differences could be observed for hsa-miR-486-5p and hsa-miR-451a, which are both the most expressed miRNAs in Illumina and CoolMPS and differ by a fold change of 76 and 87, respectively (Supplementary Figure S6A). For the Illumina samples, thus only 9.48% of the reads could be mapped to other miRNAs and after excluding the top 5 miRNAs, only 3.15% of the reads mapped to others (Supplementary Figure S6B). For the CoolMPS samples, we observed slightly increased mapping rates to the top five miRNAs on this subset of samples, with 5.60% of the reads mapping to the other miRNAs (Supplementary Figure S6C). Supplementary Figure S6D and E show a detailed breakdown of the top expressed miRNAs, after excluding the most abundant one. We also found that some miRNAs that were detected with low abundance in one technology (e.g. hsa-miR-223-3p and hsa-miR-185-5p for Illumina and hsa-miR-142-5p for CoolMPS) were among the 10 most expressed miRNAs in the other. This reinforces the necessity of deep sequencing, especially for the Illumina libraries, to quantify a larger range of miRNAs.

RT-qPCR data largely fit to the CoolMPS measurements

Finally, it is important to understand whether a third and independent technology validates the biomarker profiles. Since we previously already validated the BGISEQ approach using RT-qPCR (23) and demonstrate in the present work that CoolMPS is concordant to BGISEQ we can speculate that the RT-qPCR data would also match the CoolMPS profiles. To evaluate this hypothesis, we compared the expression values of 19 miRNAs that have been measured for 189 samples from the present study by RT-qPCR (25). Between the mean $\log_2$ CoolMPS expression and the negative delta CT values computed from RT-qPCR we observed a high correlation of 0.823 (Figure 4F). To validate how well this translates into biomarker patterns we again computed the difference between controls and dementia patients (Figure 4G). In this comparison we observed 10 miRNAs that were upregulated in both technologies, 5 miRNAs that were downregulated in both technologies and four miRNAs that were discordantly regulated between the technologies. According to Fishers Exact test this corresponds to a significant overlap ($P = 0.022$).

Table 1. For the top 10 most significant miRNAs with both technologies the rank in each technology is provided, followed by nominal and adjusted *P*-value, the effect size (Cohen's D) and AUC

miRNA	Rank BGISEQ	Rank CoolMPS	WMW raw *P*-value	WMW adj *P*-value	Cohen's D	AUC
hsa-miR-3688-3p	1	16	2.89E-06	0.0006	− 0.88	0.26
hsa-miR-3200-3p	**2**	**4**	**3.23E-06**	**0.0006**	**− 0.87**	**0.26**
hsa-let-7d-5p	3	17	6.98E-06	0.0007	0.83	0.73
hsa-miR-589-5p	4	51	9.26E-06	0.0007	− 0.79	0.27
hsa-miR-550a-3-5p	5	NA	9.39E-06	0.0007	0.78	0.73
hsa-let-7e-5p	**6**	**6**	**1.06E-05**	**0.0007**	**0.82**	**0.73**
hsa-miR-193a-3p	7	69	1.21E-05	0.0007	− 0.76	0.27
hsa-miR-4448	8	55	2.21E-05	0.0010	0.57	0.72
hsa-miR-15b-5p	**9**	**8**	**2.55E-05**	**0.0010**	**0.77**	**0.72**
hsa-miR-19b-3p	**10**	**1**	**2.64E-05**	**0.0010**	**− 0.77**	**0.28**
hsa-miR-181c-5p	21	2	2.38E-06	0.0004	− 0.80	0.26
hsa-miR-185-5p	48	3	4.84E-06	0.0006	− 0.75	0.26
hsa-miR-5695	12	5	7.70E-06	0.0006	− 0.76	0.27
hsa-miR-363-3p	33	7	2.15E-05	0.0011	0.75	0.72
hsa-miR-335-5p	117	9	5.55E-05	0.0022	− 0.44	0.29
hsa-miR-30b-5p	83	10	5.83E-05	0.0022	− 0.72	0.29

Bold miRNAs are in the top 10 for both technologies.

BGISEQ and CoolMPS AD miRNAs are matching known AD miRNAs and correlated to functional categories

As described in the previous sections, the miRNAs identified by the CoolMPS and BGISEQ approach have a significant diagnostic potential from a statistical perspective. We asked whether the signatures matched previously published results and which functional categories are enriched. To this end, we employed a miRNA set enrichment analysis using miEAA (37,38). As input the miRNAs were sorted with respect to their CoolMPS effect sizes. Downregulated miRNAs were most significantly associated to the miEAA disease category 'Downregulated in Alzheimer's Disease' (raw and adjusted *P*-value of 2.3×10^{-5} and 6.88×10^{-4}) while upregulated miRNAs were most strongly correlated to glioma (raw and adjusted *P*-value of 0.002 and 0.025, respectively). With respect to Gene Ontology and pathway databases we computed two significant categories. Upregulated AD miRNAs were enriched in chromosome condensation (raw and adjusted *P*-value of 3.3×10^{-6} and 0.018) as well as response to magnesium ion (raw and adjusted *P*-value of 1.3×10^{-5} and 0.036).

DISCUSSION

Whenever new technologies emerge in a field it is mandatory to test the fit to former technologies. The more disruptive a technological change is, the more the results differ from previous ones. An extreme example is the step from microarrays to RNA sequencing for analyzing expression profiles. If a novel technology aims to improve a previous one in a rather evolutionary manner by adapting and improving a specific step, the research results should generally be more aligned with previous findings. In biomedicine, such improvements can aim at an improved translational aspect of research in making workflows easier to use or in reducing the cost of assays. With CoolMPS we evaluated such an evolutionary improvement. Still, the main principle is sequencing-by-synthesis and also the detection and evaluation approach stay the same. The main difference is in using labeled antibodies instead of incorporating labeled nucleotides. While theoretical advantages of this approach, e.g. a potential re-use of the sequencing chemistry, are obvious we don't expect disruptive new findings. It is essential to benchmark CoolMPS to related high-throughput approaches, in our case standard cPAS sequencing-by-synthesis and Illumina sequencing, but also to a gold standard technology, in our case RT-qPCR. As primary comparison high-throughput technology we selected cPAS on the BGISEQ since we already previously performed a detailed benchmarking to the Illumina sequencing-by-synthesis approach, highlighting the advantages and disadvantages of both approaches (23). As biospecimens we intentionally selected whole blood. Not only because whole blood samples can be used to screen for minimally invasive biomarkers but also because of their challenging characteristics. The repertoire of small non-coding RNAs varies between different blood cell types and sncRNAs have a very high dynamic range. In fact, this means that few high abundant molecules are sequenced often whereas low abundant molecules are hardly observed. In whole blood small non-coding RNA sequencing data generated by Illumina sequencers, partially over 90% of the reads belong to miR-486-5p. While this miRNA is certainly highly abundant in red blood cells, this extreme distribution does not seem to match reality. In both, the BGISEQ and CoolMPS data we still observe an extreme distribution with around $\frac{3}{4}$ of all reads matching to the most abundant miRNA, miR-451a. This can also be recognized in Supplementary Figure S1K and L. Still, this distribution is less extreme than for the previously investigated Illumina sequencing data. The less extreme overrepresentation in the BGISEQ and CoolMPS data thus facilitates the discovery of yet unknown and less abundant non-coding RNA molecules.

Among the top 10 markers that we discovered by CoolMPS (Table 1), eight miRNAs (miR-19b-3p, miR-181c-5p, miR-185-5p, miR-3200-3p, let-7e-5p, miR-15b-5p, miR-335-5p and miR-30b-5p) were already described in the literature to be correlated to Alzheimer's disease or dementia. For example, miR-19b-3p prevents amyloid β-induced injury by targeting BACE1 in SH-SY5Y cells (40) and is altered in CSF exosomes of AD patients (41). Similarly, miR-185-5p is known as exosomal AD biomarker (42). Also, let-7e-5p and miR-3200-3p were previously identified as blood biomarkers (43). Interestingly, the same manuscript also lists miR-30c-5p, miR-30d-5p and miR-15a-5p. For these miRNAs we report differential expression in related miRNA family members (miR-30b-5p and miR15b-5p respectively). The latter miRNA has also been reported in other studies a circulating AD biomarker (44,45) and targets the amyloid precursor protein (46). Similarly, miR-335-5p inhibits β-Amyloid in AD (47). Already for the 10 most significant miRNAs we thus found substantial evidence for their role in AD, both as biomarker but also linked to a potential pathogenic function.

One step in our analysis pipeline is to filter out short reads (below 17 nucleotides), that might add noise to the data. For BGISEQ, the lowest number of reads was filtered out in this step followed by CoolMPS and lllumina sequencing data. While the percentages overall were similar, we observed a higher standard deviation in Illumina data (3.24%) as compared to BGISEQ (0.54) and CoolMPS (0.56) data. In comparing CoolMPS data to Illumina data we observed a slightly better averaged Q30 value for the Illumina data. This advantage could be observed however mostly in the beginning of the read. Toward the end of the 50 base reads, Illumina Q30 values dropped more as compared to the stable performance of CoolMPS. This resulted in a higher mapping rate of the CoolMPS data. One explanation for a drop of quality is in the small size of miRNAs that are usually shorter than 25 nucleotides but 50 bases are sequenced. This effect might be more pronounced for Illumina as compared to the BGISEQ and CoolMPS data. In consequence, we can expect that this factor is likely less relevant for longer RNAs or sequencing of DNA. Also, the composition of the RNA classes was different between the technologies. Illumina data revealed higher percentages of piRNAs and miRNAs while CoolMPS shows a higher diversity also including other non-coding RNA classes. A difference between the BGISEQ/CoolMPS and Illumina protocols was the amount of starting material. For BGISEQ and CoolMPS, 800 ng was used while the Illumina data have been generated from 200 ng input material. This might pretend that a higher input amount is required for CoolMPS as compared to Illumina. We used this higher input amount however only during the exploratory phase of the CoolMPS protocol. Even with lower amount of input material down to 100ng we did not observe significant changes (data not shown). Indeed, the manufacturer's instruction would even allow input from 10ng RNA only. Thus, the input volume seems not to be a limiting factor for the CoolMPS technology.

In sum, both of the technologies have their advantages and disadvantages and the best systems should be chosen dependent on the application. Our data thus clearly suggest that small RNA sequencing results from Illumina data should not be directly compared to sequencing results from BGISEQ since the technical differences between identical samples are statistically highly significant. With respect to comparing between BGISEQ and CoolMPS datasets we observed generally very similar performance. The most striking advantage of CoolMPS is a significantly improved single base call quality. This led to marginal improvements in the biomarker patterns but did not improve the performance of any biomarker in a substantial manner. Interpreting the results, we have to bear in mind that the BGISEQ technology and chemistry have already matured over at least five years while we used prototype beta testing chemistry for CoolMPS. Since already this chemistry lead to improved performance we can expect further improvements with revised kits of CoolMPS. Finally, one big advantage is the potential to recover the used labeled antibodies for a second sequencing run.

DATA AVAILABILITY

All sequencing data have been deposited in the Sequence Read Archive with the accession SRP271972.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate the support of all participants contributing to the research in providing blood samples.

FUNDING

Saarland University; Michael J. Fox Foundation for Parkinson's Research [14446]; Complete Genomics. Funding for open access charge: Michael J. Fox Foundation.

Conflict of interest statement. The authors with affiliations (2, 3, 5, 6: Y.L., A.B., S.D., S.L., C.X., M.C., C.V., A.J., H.J. and R.D.) are employed by BGI and MGI group, manufacturing the CoolMPS kits and distributing these kits commercially.

This paper is linked to: doi:10.1093/nar/gkaa1127.

REFERENCES

1. Ronaghi,M., Karamohamed,S., Pettersson,B., Uhlen,M. and Nyren,P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, **242**, 84–89.
2. Koboldt,D.C., Steinberg,K.M., Larson,D.E., Wilson,R.K. and Mardis,E.R. (2013) The next-generation sequencing revolution and its impact on genomics. *Cell*, **155**, 27–38.
3. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
4. Saliba,A.E., Westermann,A.J., Gorski,S.A. and Vogel,J. (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.*, **42**, 8845–8860.
5. Slatko,B.E., Gardner,A.F. and Ausubel,F.M. (2018) Overview of next-generation sequencing technologies. *Curr. Protoc. Mol. Biol.*, **122**, e59.
6. Senabouth,A., Andersen,S., Shi,Q., Shi,L., Jiang,F., Zhang,W., Wing,K., Daniszewski,M., Lukowski,S.W., Hung,S.S.C. *et al.* (2020) Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genomics Bioinform.*, **2**, lqaa034.

7. Mathew,R., Mattei,V., Al Hashmi,M. and Tomei,S. (2020) Updates on the current technologies for microRNA profiling. *Microrna*, **9**, 17–24.
8. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grasser,F.A., Lenhof,H.P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, **47**, 3353–3364.
9. Fehlmann,T., Backes,C., Alles,J., Fischer,U., Hart,M., Kern,F., Langseth,H., Rounge,T., Umu,S.U., Kahraman,M. *et al.* (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, **34**, 1621–1628.
10. Fehlmann,T., Backes,C., Pirritano,M., Laufer,T., Galata,V., Kern,F., Kahraman,M., Gasparoni,G., Ludwig,N., Lenhof,H.P. *et al.* (2019) The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Res.*, **47**, 4431–4441.
11. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
12. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
13. Fromm,B., Domanska,D., Hoye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2020) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, **48**, D132–D141.
14. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.P., Meese,E. and Keller,A. (2018) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
15. Fromm,B., Keller,A., Yang,X., Friedlander,M.R., Peterson,K.J. and Griffiths-Jones,S. (2020) Quo vadis microRNAs? *Trends Genet*, **36**, 461–463.
16. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Wurstle,M.L., Hubenthal,M., Franke,A., Meder,B *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.
17. Friedlander,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.
18. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
19. Heinicke,F., Zhong,X., Zucknick,M., Breidenbach,J., Sundaram,A.Y.M., S,T.F., Leithaug,M., Dalland,M., Farmer,A., Henderson,J.M. *et al.* (2020) Systematic assessment of commercially available low-input miRNA library preparation kits. *RNA Biol.*, **17**, 75–86.
20. Meistertzheim,M., Fehlmann,T., Drews,F., Pirritano,M., Gasparoni,G., Keller,A. and Simon,M. (2019) Comparative analysis of biochemical biases by ligation- and template-switch-Based small RNA library preparation protocols. *Clin. Chem.*, **65**, 1581–1591.
21. Ludwig,N., Fehlmann,T., Galata,V., Franke,A., Backes,C., Meese,E. and Keller,A. (2018) Small ncRNA-Seq results of human Tissues: Variations depending on sample integrity. *Clin. Chem.*, **64**, 1074–1084.
22. Baroin-Tourancheau,A., Jaszczyszyn,Y., Benigni,X. and Amar,L. (2019) Evaluating and correcting inherent bias of microRNA expression in Illumina sequencing analysis. *Front. Mol. Biosci.*, **6**, 17.
23. Fehlmann,T., Reinheimer,S., Geng,C., Su,X., Drmanac,S., Alexeev,A., Zhang,C., Backes,C., Ludwig,N., Hart,M. *et al.* (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet.*, **8**, 123.
24. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement*, **12**, 565–576.
25. Ludwig,N., Fehlmann,T., Kern,F., Gogol,M., Maetzler,W., Deutscher,S., Gurlit,S., Schulte,C., von Thaler,A.K., Deuschle,C. *et al.* (2019) Machine learning to detect Alzheimer's disease from circulating Non-coding RNAs. *Genomics Proteomics Bioinform.*, **17**, 430–440.
26. Fehlmann,T., Meese,E. and Keller,A. (2017) Exploring ncRNAs in Alzheimer's disease by miRMaster. *Oncotarget*, **8**, 3771–3772.
27. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
28. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
29. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
30. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
31. Wang,J., Zhang,P., Lu,Y., Li,Y., Zheng,Y., Kan,Y., Chen,R. and He,S. (2019) piRBase: a comprehensive database of piRNA sequences. *Nucleic Acids Res.*, **47**, D175–D180.
32. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.
33. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grasser,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.
34. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.C. and Muller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.
35. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
36. Amand,J., Fehlmann,T., Backes,C. and Keller,A. (2019) DynaVenn: web-based computation of the most significant overlap between ordered sets. *BMC Bioinformatics*, **20**, 743.
37. Kern,F., Fehlmann,T., Solomon,J., Schwed,L., Grammes,N., Backes,C., Van Keuren-Jensen,K., Craig,D.W., Meese,E. and Keller,A. (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res.*, **48**, W521–W528.
38. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
39. Amrhein,V., Greenland,S. and McShane,B. (2019) Scientists rise up against statistical significance. *Nature*, **567**, 305–307.
40. Zhang,N., Li,W.W., Lv,C.M., Gao,Y.W., Liu,X.L. and Zhao,L. (2020) miR-16-5p and miR-19b-3p prevent amyloid beta-induced injury by targeting BACE1 in SH-SY5Y cells. *Neuroreport*, **31**, 205–212.
41. Gui,Y., Liu,H., Zhang,L., Lv,W. and Hu,X. (2015) Altered microRNA profiles in cerebrospinal fluid exosome in Parkinson disease and Alzheimer disease. *Oncotarget*, **6**, 37043–37053.
42. Lugli,G., Cohen,A.M., Bennett,D.A., Shah,R.C., Fields,C.J., Hernandez,A.G. and Smalheiser,N.R. (2015) Plasma exosomal miRNAs in persons with and without Alzheimer Disease: Altered expression and prospects for biomarkers. *PLoS One*, **10**, e0139233.
43. Satoh,J., Kino,Y. and Niida,S. (2015) MicroRNA-seq data analysis pipeline to identify blood biomarkers for Alzheimer's disease from public data. *Biomark Insights*, **10**, 21–31.
44. Kumar,P., Dezso,Z., MacKenzie,C., Oestreicher,J., Agoulnik,S., Byrne,M., Bernier,F., Yanagimachi,M., Aoshima,K. and Oda,Y. (2013) Circulating miRNA biomarkers for Alzheimer's disease. *PLoS One*, **8**, e69807.
45. Wu,H.Z.Y., Thalamuthu,A., Cheng,L., Fowler,C., Masters,C.L., Sachdev,P., Mather,K.A. and and the Australian Imaging, B. and Lifestyle Flagship Study of, A. (2020) Differential blood miRNA expression in brain amyloid imaging-defined Alzheimer's disease and controls. *Alzheimers Res. Ther.*, **12**, 59.
46. Liu,H.Y., Fu,X., Li,Y.F., Li,X.L., Ma,Z.Y., Zhang,Y. and Gao,Q.C. (2019) miR-15b-5p targeting amyloid precursor protein is involved in the anti-amyloid eflect of curcumin in swAPP695-HEK293 cells. *Neural. Regen. Res.*, **14**, 1603–1609.
47. Wang,D., Fei,Z., Luo,S. and Wang,H. (2020) MiR-335-5p Inhibits beta-Amyloid (Abeta) accumulation to attenuate cognitive deficits through targeting c-jun-N-terminal kinase 3 in Alzheimer's disease. *Curr. Neurovasc. Res.*, **17**, 93–101.

3.18 MiRTargetLink-miRNAs, genes and interaction networks

International Journal of *Molecular Sciences*

MDPI

Communication

miRTargetLink—miRNAs, Genes and Interaction Networks

Maarten Hamberg [1], Christina Backes [1], Tobias Fehlmann [1], Martin Hart [2], Benjamin Meder [3], Eckart Meese [2] and Andreas Keller [1,*]

[1] Chair for Clinical Bioinformatics, Saarland University, Saarbrücken D-66041, Germany; e.j.m.hamberg@gmail.com (M.Ham.); c.backes@mx.uni-saarland.de (C.B.); tobias.fehlmann@ccb.uni-saarland.de (T.F.)

[2] Department of Human Genetics, Saarland University, Homburg D-66421, Germany; marhar15@googlemail.com (M.Har.); hgemee@uks.eu (E.M.)

[3] Internal Medicine, Heidelberg University, Heidelberg D-69120, Germany; benjamin.meder@med.uni-heidelberg.de

* Correspondence: andreas.keller@ccb.uni-saarland.de; Tel.: +49-174-168-4638

Academic Editor: Y-h. Taguchi
Received: 25 March 2016; Accepted: 7 April 2016; Published: 14 April 2016

Abstract: Information on miRNA targeting genes is growing rapidly. For high-throughput experiments, but also for targeted analyses of few genes or miRNAs, easy analysis with concise representation of results facilitates the work of life scientists. We developed miRTargetLink, a tool for automating respective analysis procedures that are frequently applied. Input of the web-based solution is either a single gene or single miRNA, but also sets of genes or miRNAs, can be entered. Validated and predicted targets are extracted from databases and an interaction network is presented. Users can select whether predicted targets, experimentally validated targets with strong or weak evidence, or combinations of those are considered. Central genes or miRNAs are highlighted and users can navigate through the network interactively. To discover the most relevant biochemical processes influenced by the target network, gene set analysis and miRNA set analysis are integrated. As a showcase for miRTargetLink, we analyze targets of five cardiac miRNAs. miRTargetLink is freely available without restrictions at www.ccb.uni-saarland.de/mirtargetlink.

Keywords: miRTargetkLink; miRNAs; genes; interaction networks

1. Introduction

Biochemical networks play a central role in life science research. Tools for visualizing and analyzing such networks have a long history. Among the most popular web-based applications is STRING. Originally implemented as a search tool for recurring instances of neighboring genes [1], STRING v10 has become a comprehensive protein-protein interaction database containing predicted and validated protein-protein target information [2]. The success of this web service and database is not only driven by the large amount of information in the database but also the intuitive manner in its running analyses. Entering one or several protein identifiers enables comprehensive analysis of protein-protein interactions and deriving important information on a network level.

With non-coding RNAs gaining importance in biomedical research, similar tools with respect to miRNA-gene interactions have been presented. We here exemplarily mention three web services that involve the analysis of miRNAs and target genes. Among the first tools, we published miRTrail [3]. Tailored for high-throughput omics analyses, deregulated genes, miRNAs and target information can be uploaded to find clusters. Hamed and co-workers presented TFmiR, a web service for constructing and analyzing transcription factor and miRNA co-regulatory networks in a disease-specific manner [4].

One further tool in a similar direction is TargetCompare, a web interface for studying multiple targets of pre-selected miRNAs [5].

Generally, web services of miRNA-gene interactions require file data upload (e.g., gene or miRNA expression) and are tailored for specific applications. The majority of available tools are designated predominantly for large-scale experimental data such as microarrays and high-throughput sequencing. Many researchers work, however, with single genes or miRNAs or, at most, with small sets. If regulatory information is included in respective research projects, targets are frequently extracted manually from databases and small networks are drawn manually or using tools such as Cytoscape [6]. Taking STRING as an archetype, we implemented miRTargetLink. Users can enter single miRNAs, genes or sets of miRNAs and genes (ranging from two or three up to very comprehensive sets from high-throughput experiments) in the web interface. Interaction networks based on the information from miRTarBase [7] and miRanda [8] are calculated and visualized. Users can modify the resulting network, e.g., only validated targets with strong evidence can be considered. For downstream analysis we have built an interface to GeneTrail2, an updated version of the GeneTrail web service [9].

In this manuscript we describe miRTargetLink, as well as its input, output and the functionality. We also mention current limitations and forward research directions. On the webpage of miRTargetLink [10], example input is provided and alongside other relevant information, a detailed tutorial is also available.

2. Results

One of the major goals of miRTargetLink, besides functionality, was a user-friendly interface. To provide a convenient and standardized solution that runs on state-of-the art web browsers, miRTargetLink is implemented in php 5 and JavaScript. Interactions are stored in a MySQl database that is regularly updated. The interaction networks are generated with the networks visualization package from the VIS.js JavaScript library version 4 [11].

3. Data Input

An important feature of miRTargetLink is the straightforward data input. Single miRNA identifiers, gene identifiers or sets of miRNAs and genes can be copy-pasted in the web interface. No file data upload is required. For gene identifiers, the gene symbol is used, and miRNAs are uploaded following the current nomenclature that is also used by miRBase [12]. Since many research projects have been carried out on information from older miRBase versions (using the * annotation for minor forms of miRNAs), we implemented a web-based mapping tool that converts identifiers from older versions to the most recent annotation [13]. The converted IDs can then be provided to miRTargetLink. Gene and miRNA identifiers are matched automatically; if several hits are found for a provided ID, the user can select the right hit. If, for example, "let-7a" is entered, the three matching hits hsa-let-7a-2-3p, hsa-let-7a-3p and hsa-let-7a-5p are proposed by miRTargetLink. In the single gene/miRNA mode, the correct ID is selected by a radio button. In the multi-miRNA/gene mode, for each gene/miRNA the number of hits is shown and the correct ID can be selected from a drop-down list.

4. Constructing, Visualizing and Modifying Networks

From the input, target information is extracted from miRTarBase [7], one of the most comprehensive miRNA-gene target resources presently available. Predicted targets are added based on the miRanda algorithm. From the data, an interaction network is calculated and visualized.

As an example, for the single miRNA mode the network for miR-107 is presented in Figure 1, where the central brown node is the miRNA, the green nodes in the inner circle are validated targets with strong evidence (e.g., luciferase assay), the blue genes in the middle circle are validated by weak evidence (e.g., microarray), and the yellow nodes in the outer circle are the predicted targets. The second image is the result of removing the predicted targets. The representation in the lower left part results from removing weak experimental evidence targets as well. By double-clicking a gene node, e.g., *CAV1*, the network-centric view for this gene is generated (lower right part). Now the miRNAs that are known to or predicted to target *CAV1* are presented in the respective colors. In each case, the current interactions of the network are dynamically listed in tabular form below the visualization widget. The respective table can also be downloaded as a text file. The genes and miRNAs in that table are linked to GeneCards and miRBase. The network visualization can be saved as a png file using the right mouse button.

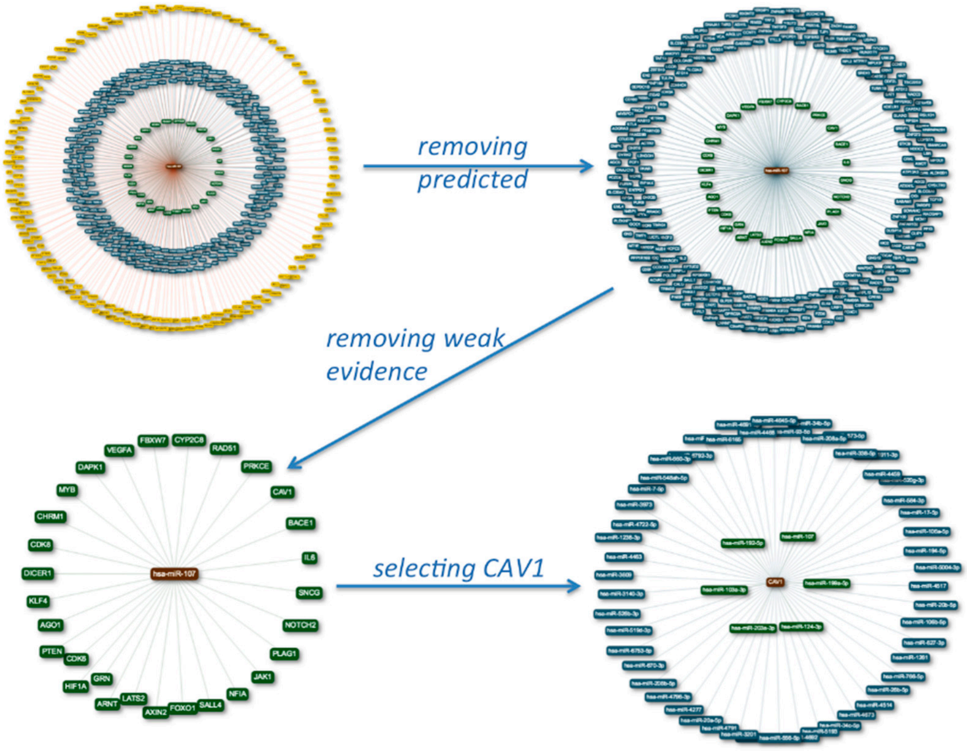


Figure 1. Single miRNA mode for miR-107: The central node is the miRNA, surrounded by validated targets with strong evidence (green), weak evidence (blue), and predicted targets (yellow). In the second representation predicted edges and in the third weak evidence edges are removed. The final representation is gene-centric for one of the target gens of miR-107, *CAV1*.

Often, researchers are, however, not only interested in single genes or miRNAs but in sets of the respective molecules. Thus, miRTargetLink can also handle comma-separated lists of identifiers, both for genes and miRNAs. As a case study, we used five known cardio-miRNAs as input, including hsa-miR-1-3p [14,15], hsa-miR-145-5p [16,17], hsa-miR-30a-5p [17], hsa-miR-30c-5p [17], hsa-miR-423-5p [15]. Users can select the interactions to be included in the network, *i.e.*, weak or strong evidence-validated interactions. The network containing all validated targets for the five cardio miRNAs is presented in Figure 2. The color of the genes indicates the number of interactions. Orange genes are targeted by three or more microRNAs, genes that are targeted by two microRNAs are blue. Interactions with strong experimental evidence are depicted by green edges while interactions with weaker evidence are depicted by blue (genes with two edges) or orange (genes with more than two edges). As for the single gene/miRNA mode, weak evidence edges can be removed. The resulting network for the same miRNAs is presented in the right part of Figure 2. Genes or miRNAs can be marked from this network and novel sub-networks can be generated from the marked nodes in a new window, making the multi-gene/miRNA mode as interactive as the single-gene mode.

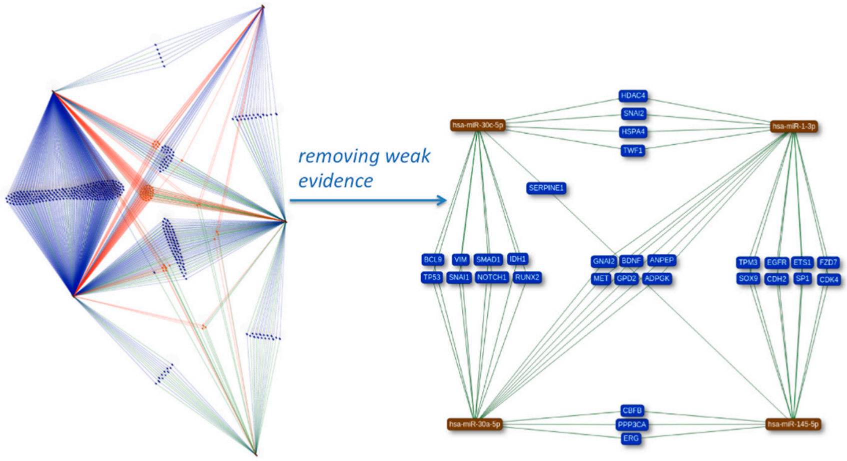


Figure 2. Multi-miRNA mode. For five cardio miRNAs, the network of genes that are targeted by at least two miRNAs (experimental evidence) is shown. Edge color indicates the number of miRNA-target interactions for the miRNAs that are start nodes of the edges. The second representation highlights the genes that are supported by strong evidence experiments.

5. Enrichment Analysis

As result of the target analysis, sets of miRNAs and genes are generated dynamically. An important aspect for researchers is the enrichment of genes or miRNAs in biochemically relevant categories such as pathways. We implemented an interface to GeneTrail2 [18], which builds up on the GeneTrail framework [9]. GeneTrail2 provides the respective functionality for systems biological analysis of the sets and enables over-representation analysis with a single mouse click. As a result, gene ontology categories, KEGG pathways, transcription factors or genomic clusters that are enriched for genes in the network are presented. Although single miRNAs can potentially regulate pathways [19], this analysis functionality is tailored for sets of miRNAs, e.g., in the case of the five cardio miRNAs, we find enrichment for cancer-related miRNAs, driven by markers that are enriched in the cell cycle and the apoptosis. Among the most significant KEGG pathways we observe "Arrhythmogenic right ventricular cardiomyopathy" (Benjamini-Hochberg adjusted p-value of 1.6×10^{-4}) with genes *JUP*, *ACTB*, *ITGA6*, *DSG2*, *ATP2A2*, *ITGB4*, *DAG1* and *CDH2*. Also "Dilated cardiomyopathy" with genes *ACTB*, *ITGA6*, *ATP2A2*, *ITGB4*, *DAG1* and *TPM3* remained significant (Benjamini-Hochberg adjusted p-value of 6.6×10^{-3}).

6. Limitations and Future Research Directions

We consider miRTargetLink as an ongoing research project. Thus, we want to mention current limitations of the tool and three ongoing research efforts.

(1) At this stage, the application is limited to *Homo sapiens* since most information is available and a considerable amount of biomedical research is carried out in humans. In the next release we will integrate other organisms, starting with *Mus musculus* and *Rattus norvegicus*. Along with more organisms, improved mapping functionality and support of other identifiers can facilitate the application of miRTargetLink.

(2) Validated miRNA gene interactions are central for miRTargetLink. As a comprehensive resource for such interactions, we selected the miRTarBase. In addition to this database, many others are available (e.g., Tarbase or miRecords). To extend the tool beyond validated targets, predicted interactions are included. However, respective targets are known to depend on the prediction algorithm. Results can vary tremendously between different algorithms. In the present release, predicted interactions in miRTargetLink rely on miRanda. We have selected miRanda as a case of

frequently applied prediction algorithms. Adding other prediction tools can potentially improve the results provided by miRTargetLink. Before adding those, the performance of the different methods, however, has to be evaluated critically to validate interactions.

(3) The miRNA-gene predictions are known to be error-prone. This means that some of the miRNA-gene interactions that have been validated by life scientists may have been negative. The negative experiments are usually not reported. We are implementing a database for negative miRNA-gene interactions and will integrate the information in miRTargetLink so that potential users also get the information on false positive predictions.

7. Discussion

With miRTargetLink we present a web interface facilitating low- and medium-throughput analysis of miRNAs and target genes. We mainly address applied life scientist researchers who work with single genes, miRNAs or small sets to facilitate tasks that are currently done on a daily basis but in a semi-automated fashion.

While miRTargetLink is not thought to be a replacement for existing high-throughput miRNA analytics tools that also include quantitative analysis (such as miRTrail [3], TFmiR [4] or TargetCompare [5]), our program can also be used for quantitative high-throughput miRNA and gene expression analyses, e.g., from microarrays or Next Generation Sequencing, we implemented straightforward data input and handling. Additional information along with a detailed tutorial is provided online. Currently implemented for *Homo sapiens*, we consider miRTargetLink an ongoing research project. Extension to other model organisms is a reasonable next step, together with additional background information such as false positive miRNA-gene interactions that have been proven to be wrong in functional experiments.

Acknowledgments: Funding: This work was supported by the EU FP7 project BestAgeing.

Author Contributions: Maarten Hamberg and Andreas Keller conceived and designed miRTargetLink; Maarten Hamberg implemented the web-service; Andreas Keller performed the case study; Benjamin Meder contributed to the case study; Andreas Keller wrote the paper; Christina Backes, Tobias Fehlmann, Martin Hart and Eckart Meese supported the development, proof read the manuscript and tested the system.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Snel, B.; Lehmann, G.; Bork, P.; Huynen, M.A. String: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **2000**, *28*, 3442–3444. [CrossRef] [PubMed]
2. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P.; *et al.* String v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–D452. [CrossRef] [PubMed]
3. Laczny, C.; Leidinger, P.; Haas, J.; Ludwig, N.; Backes, C.; Gerasch, A.; Kaufmann, M.; Vogel, B.; Katus, H.A.; Meder, B.; *et al.* miRTrail—A comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. *BMC Bioinform.* **2012**, *13*. [CrossRef] [PubMed]
4. Hamed, M.; Spaniol, C.; Nazarieh, M.; Helms, V. TFmiR: A web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks. *Nucleic Acids Res.* **2015**, *43*, W283–W288. [CrossRef] [PubMed]
5. Moreira, F.C.; Dustan, B.; Hamoy, I.G.; Ribeiro-Dos-Santos, A.M.; Dos Santos, A.R. TargetCompare: A web interface to compare simultaneous miRNAs targets. *Bioinformation* **2014**, *10*, 602–605. [CrossRef] [PubMed]
6. Smoot, M.E.; Ono, K.; Ruscheinski, J.; Wang, P.L.; Ideker, T. Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* **2011**, *27*, 431–432. [CrossRef] [PubMed]
7. Chou, C.H.; Chang, N.W.; Shrestha, S.; Hsu, S.D.; Lin, Y.L.; Lee, W.H.; Yang, C.D.; Hong, H.C.; Wei, T.Y.; Tu, S.J.; *et al.* miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **2016**, *44*, D239–D247. [CrossRef] [PubMed]

8. Enright, A.J.; John, B.; Gaul, U.; Tuschl, T.; Sander, C.; Marks, D.S. MicroRNA targets in Drosophila. *Genome Biol.* **2003**, *5*. [CrossRef] [PubMed]
9. Backes, C.; Keller, A.; Kuentzer, J.; Kneissl, B.; Comtesse, N.; Elnakady, Y.A.; Muller, R.; Meese, E.; Lenhof, H.P. GeneTrail—Advanced gene set enrichment analysis. *Nucleic Acids Res.* **2007**, *35*, W186–W192. [CrossRef] [PubMed]
10. miRTargetLink. Available online: http://ccb-web.cs.uni-saarland.de/mirtargetlink (accessed on 8 April 2016).
11. VIS.js. Available online: http://www.visjs.org (accessed on 8 April 2016).
12. Kozomara, A.; Griffiths-Jones, S. miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **2014**, *42*, D68–D73. [CrossRef] [PubMed]
13. miEAA. Available online: www.ccb.uni-saarland.de/mieaa_tool/mirna_version_converter (accessed on 8 April 2016).
14. Duan, L.; Xiong, X.; Liu, Y.; Wang, J. miRNA-1: Functional roles and dysregulation in heart disease. *Mol. Biosyst.* **2014**, *10*, 2775–2782. [CrossRef] [PubMed]
15. Nabialek, E.; Wanha, W.; Kula, D.; Jadczyk, T.; Krajewska, M.; Kowalowka, A.; Dworowy, S.; Hrycek, E.; Wludarczyk, W.; Parma, Z.; *et al.* Circulating microRNAs (miR-423-5p, miR-208a and miR-1) in acute myocardial infarction and stable coronary heart disease. *Minerva Cardioangiol.* **2013**, *61*, 627–637. [PubMed]
16. Higashi, K.; Yamada, Y.; Minatoguchi, S.; Baba, S.; Iwasa, M.; Kanamori, H.; Kawasaki, M.; Nishigaki, K.; Takemura, G.; Kumazaki, M.; *et al.* MicroRNA-145 repairs infarcted myocardium by accelerating cardiomyocyte autophagy. *Am. J. Physiol. Heart Circ. Physiol.* **2015**, *309*, H1813–H1826. [CrossRef] [PubMed]
17. Liu, Q.; Du, G.Q.; Zhu, Z.T.; Zhang, C.; Sun, X.W.; Liu, J.J.; Li, X.; Wang, Y.S.; Du, W.J. Identification of apoptosis-related microRNAs and their target genes in myocardial infarction post-transplantation with skeletal myoblasts. *J. Transl. Med.* **2015**, *13*. [CrossRef] [PubMed]
18. Stockel, D.; Kehl, T.; Trampert, P.; Schneider, L.; Backes, C.; Ludwig, N.; Gerasch, A.; Kaufmann, M.; Gessler, M.; Graf, N.; *et al.* Multi-omics Enrichment Analysis using the GeneTrail2 Web Service. *Bioinformatics* **2016**, *32*. [CrossRef] [PubMed]
19. Backes, C.; Meese, E.; Lenhof, H.P.; Keller, A. A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.* **2010**, *38*, 4476–4486. [CrossRef] [PubMed]

D90–D96 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 13 October 2016*
doi: 10.1093/nar/gkw926

miRPathDB: a new dictionary on microRNAs and target pathways

Christina Backes[1,†], Tim Kehl[2,†], Daniel Stöckel[2], Tobias Fehlmann[1], Lara Schneider[2], Eckart Meese[3], Hans-Peter Lenhof[2] and Andreas Keller[1,*]

[1]Chair for Clinical Bioinformatics, Saarland Informatics Campus, Saarland University, D-66123 Saarbruecken, Germany, [2]Chair for Bioinformatics, Saarland Informatics Campus, Saarland University, D-66123 Saarbruecken, Germany and [3]Human Genetics, Saarland University, D-66421 Homburg, Germany

Received September 19, 2016; Accepted October 10, 2016

ABSTRACT

In the last decade, miRNAs and their regulatory mechanisms have been intensively studied and many tools for the analysis of miRNAs and their targets have been developed. We previously presented a dictionary on single miRNAs and their putative target pathways. Since then, the number of miRNAs has tripled and the knowledge on miRNAs and targets has grown substantially. This, along with changes in pathway resources such as KEGG, leads to an improved understanding of miRNAs, their target genes and related pathways. Here, we introduce the miRNA Pathway Dictionary Database (miRPathDB), freely accessible at https://mpd.bioinf.uni-sb.de/. With the database we aim to complement available target pathway web-servers by providing researchers easy access to the information which pathways are regulated by a miRNA, which miRNAs target a pathway and how specific these regulations are. The database contains a large number of miRNAs (2595 human miRNAs), different miRNA target sets (14 773 experimentally validated target genes as well as 19 281 predicted targets genes) and a broad selection of functional biochemical categories (KEGG-, WikiPathways-, BioCarta-, SMPDB-, PID-, Reactome pathways, functional categories from gene ontology (GO), protein families from Pfam and chromosomal locations totaling 12 875 categories). In addition to *Homo sapiens*, also *Mus musculus* data are stored and can be compared to human target pathways.

INTRODUCTION

The understanding of regulatory mechanisms of non-coding RNAs is growing rapidly. Small non-coding RNAs, so called miRNAs or microRNAs play a central role in the regulation of molecular pathways. Already in 2001, miRNAs were described as ‘tiny regulators with great potential (1). The most comprehensive collection of miRNAs is the miRBase that can be considered as reference database and central repository. First published in 2004 with 506 miRNAs from six organisms (2), the 10th release published in 2008 already contained 5071 miRNA precursors from 58 species (3). These correspond to 5922 mature miRNA sequences.

For the analysis of miRNAs a wide variety of computational tools has been developed in the last decade. Akhtar *et al.* published a comprehensive review containing the description of 129 stand-alone and web-based analysis packages (4). One important task in miRNA research is to understand which pathways are regulated either by single miRNAs or miRNA sets. A selection of tools that provide solution for this task includes miRNApath (5) a Bioconductor package for enrichment of miRNA expression data, miTALOS (6) a web-server for the analysis of tissue specific regulation in signaling pathways or DIANA-miRPath (7,8), a broad target pathway analysis tool that is regularly updated. Similar functionality is also included in general ‘omics’ enrichment toolboxes such as GeneTrail2 (9) and in miRNA analysis pipelines such as Oasis (10).

Similar to the approaches implemented in the tools mentioned above, we performed an *in silico* enrichment analysis for single miRNAs in 2010. We asked whether predicted target genes of miRNAs accumulate in certain KEGG pathways or gene ontologies and how specific the regulation is. The result was the dictionary of miRNAs and their putative target pathways (11). In the past six years, the knowledge on miRNAs has improved tremendously. The most recent version 21 of miRBase (12)—available since June 2014—lists 28 645 precursor miRNAs, expressing 35 828 mature miRNA products, in 223 different species. For *Homo sapiens* alone, 2600 miRNAs are annotated and the number has tripled compared to the data contained in our miRNA to target pathway dictionary. In addition to

[*]To whom correspondence should be addressed. Tel: +49 68130268611; Fax: +49 68130268612; Email: andreas.keller@ccb.uni-saarland.de
[†]These authors contributed equally to the study as the first authors.

miRNA resources, also pathway databases, gene ontologies and bioinformatics methods have been extended significantly in the past six years. With GeneTrail2 (9), we developed a comprehensive gene set analysis toolbox containing substantially increased functionality as compared to the original version GeneTrail (13).

Since it has become evident that *in silico* target prediction bears substantial challenges (14), increased experimental effort has led to a comprehensive collection of miRNAs and validated target genes. Among the most comprehensive databases storing miRNA-target interactions (MTIs) are TarBase (15) and miRTarBase (16). The latter contains 2599 human miRNAs and 14 773 genes targeted by at least a single miRNA. Here, MTIs are classified either as 'strong evidence' or 'weak evidence'. Strong evidence targets include interactions validated by reporter assay, Western blot and qPCR. Weak evidence interactions are based on microarrays, next-generation sequencing, pSILAC and other experiments.

The increased knowledge on miRNAs, the availability of large sets of experimentally validated miRNA targets and changes in gene set enrichment and pathway analyses call for the incorporation of this state-of-the-art information in an updated version of miRNA target pathway dictionary. For each miRNA we created three sets of experimentally validated and two sets of predicted MTIs. For the experimentally validated sets, we extracted for each miRNA all targets from miRTarBase and used them to create three test sets: (i) MTIs validated by any experimental method, (ii) MTIs validated by methods with strong evidence and (iii) MTIs validated with weak experimental evidence. In order to build the predicted MTI sets, we used three well established miRNA target prediction frameworks: DIANA-microT (24), miRDB (25) and TargetScan (26). From each of those, we extracted all precomputed MTIs and created two further sets containing the intersection and the union of the three predicted data sets. Using the five MTI sets we computed for each human miRNA enrichments based on 280 KEGG pathways (17), 1300 pathways from Reactome (18), 310 pathways from BioCarta (19), 6169 gene ontology (GO) (20) categories (molecular function, cellular components and biological processes), 617 categories from the Small Molecule Pathway Database (22), as well as enrichments based on the 14 chromosomes, 806 cytogenetic bands, 560 Pfam protein families (28) and on 221 categories from the National Cancer Institute (NCI) Pathway Interaction Database (21). In sum, 12 875 different functional categories were investigated. All results have been stored in the miRNA Pathway Dictionary Database (miRPathDB), freely accessible at https://mpd.bioinf.uni-sb.de/. Altogether, our database stores significant interactions for 2571 miRNAs and 7565 functional categories for *Homo sapiens*. Besides human, we also incorporated significant interactions for 1933 miRNAs and 8201 functional categories for *Mus musculus* and enable comparison of these to understand whether miRNA pathway regulations are conserved between organisms.

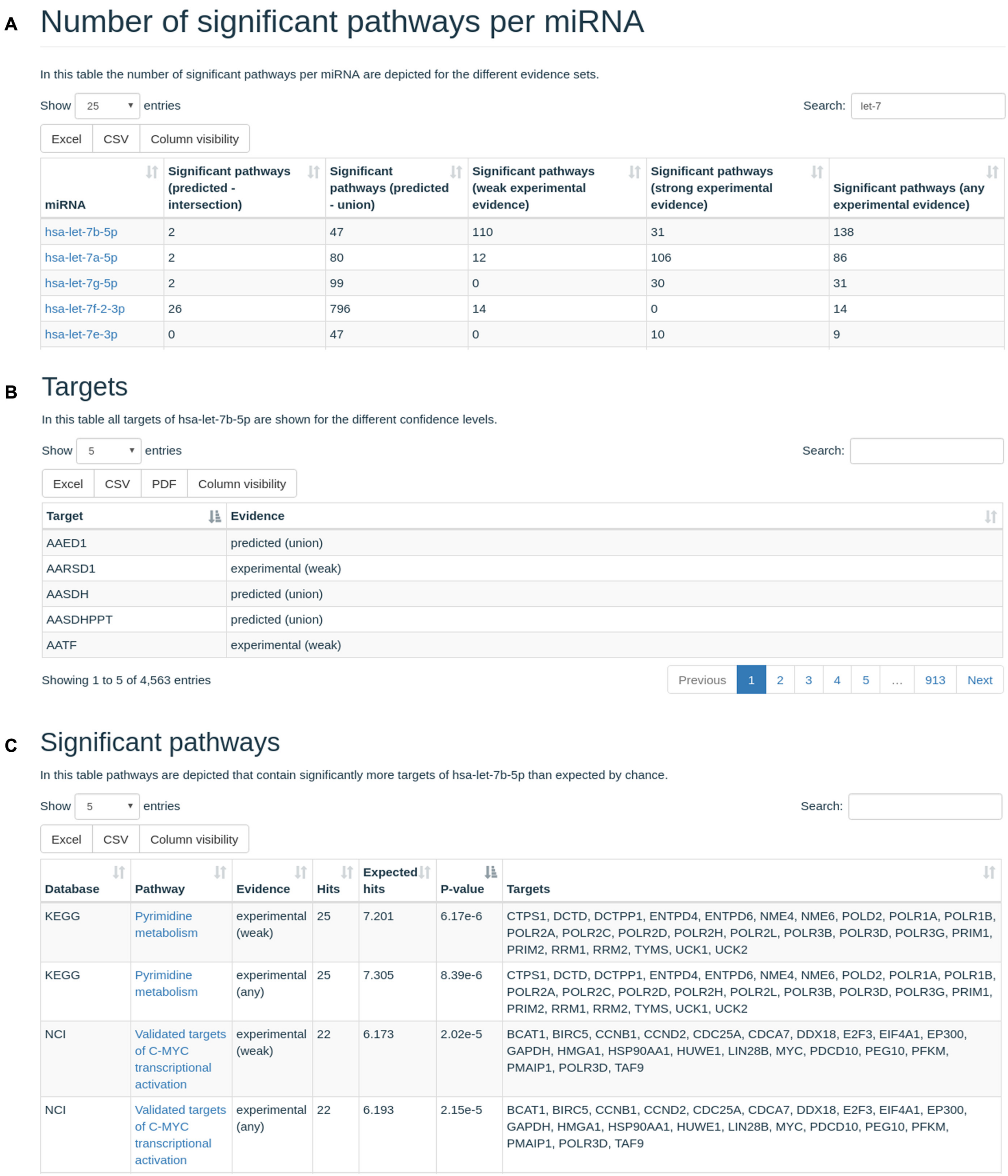
Data sources and enrichment analyses for miRPathDB

Data sources. With miRPathDB, we strive to provide a new database resource that covers a broad range of miRNAs, target genes, as well as potentially enriched pathways and functional categories. The results included in miRPathDB generally rely on three data sources: human or mouse miRNAs, targets of miRNAs and functional categories. With respect to the miRNAs, we use information from the miRBase version 21 (2). Predicted MTIs are extracted from precomputed data sets provided by DIANA-microT (24), miRDB (25) and TargetScan (26). Based on these data sets, we created two test sets containing the intersection and the union of the provided predictions for each miRNA. Experimentally validated MTIs were retrieved from miRTarBase version 6 (16). The respective MTIs were then used to create three test sets with different experimental evidence levels (any, weak and strong). The third resource, functional categories, are obtained from the GeneTrail2 (9) data warehouse. This data warehouse includes functional categories from various third party resources including KEGG (17), Reactome (18), BioCarta (19), GO (20), cytogenetic bands, disease categories from the NCI Pathway Interaction Database (21) and the Small Molecule Pathway Database (22). Altogether, we considered 12 875 functional categories for human and 9741 for mouse in the current statistical analysis.

Enrichment analysis and clustering. In order to identify if the targets of a certain miRNA are enriched in a biological category, we use the hypergeometric test implemented in the GeneTrail2 C++ library (9). This test checks if this category contains more targets of the analyzed miRNA then expected by chance. In order to calculate this chance, the hypergeometric test relies on a reference set (background). In our case, this is the list of all miRNA targets in the corresponding target sets (weak experimental evidence, strong experimental evidence, any experimental evidence, intersection of predicted targets, union of predicted targets).

For each miRNA and the associated test sets, the following analysis is carried out: We used the hypergeometric test to compute a *P*-value for the given test set, reference set and biological category. Multiple testing corrections were performed by controlling the false discovery rate (Benjamini–Hochberg adjustment). The significance level was set to 0.05. We only focused on significantly enriched categories rather than depleted ones. *P*-values for depleted categories were set to 1. The minimal category size was set to 2, the maximal category size to 1000.

To process the results of the enrichment analyses and to enable the integrative visualization as heat maps, we used the freely available statistical programming environment R, version 3.0.2. In order to build heat maps for each database and target set category, we used the following methodology. We used the GeneTrail2 enrichment results to build a pathway x miRNA *P*-value matrix for each database and target set category. The respective *P*-values were log10 transformed and discretized. We then performed complete linkage hierarchical clustering with the Euclidian distance, using the hclust method from the stats package, in order to group similar signatures.



Number of significant pathways per miRNA

In this table the number of significant pathways per miRNA are depicted for the different evidence sets.

Show 25 entries

Search: let-7

miRNA	Significant pathways (predicted - intersection)	Significant pathways (predicted - union)	Significant pathways (weak experimental evidence)	Significant pathways (strong experimental evidence)	Significant pathways (any experimental evidence)
hsa-let-7b-5p	2	47	110	31	138
hsa-let-7a-5p	2	80	12	106	86
hsa-let-7g-5p	2	99	0	30	31
hsa-let-7f-2-3p	26	796	14	0	14
hsa-let-7e-3p	0	47	0	10	9

Targets

In this table all targets of hsa-let-7b-5p are shown for the different confidence levels.

Show 5 entries

Search:

Target	Evidence
AAED1	predicted (union)
AARSD1	experimental (weak)
AASDH	predicted (union)
AASDHPPT	predicted (union)
AATF	experimental (weak)

Showing 1 to 5 of 4,563 entries

Significant pathways

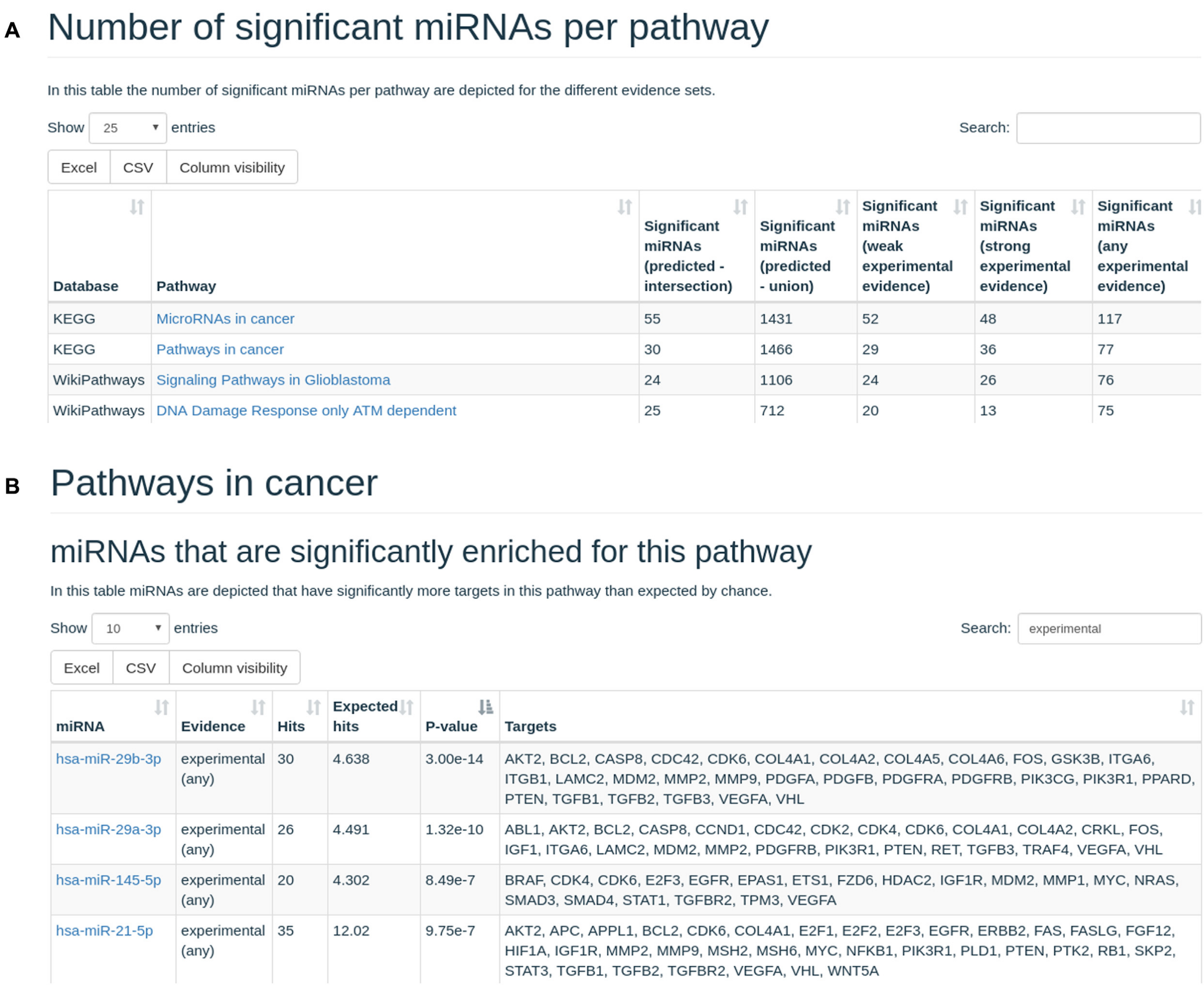
In this table pathways are depicted that contain significantly more targets of hsa-let-7b-5p than expected by chance.

Show 5 entries

Search:

Database	Pathway	Evidence	Hits	Expected hits	P-value	Targets
KEGG	Pyrimidine metabolism	experimental (weak)	25	7.201	6.17e-6	CTPS1, DCTD, DCTPP1, ENTPD4, ENTPD6, NME4, NME6, POLD2, POLR1A, POLR1B, POLR2A, POLR2C, POLR2D, POLR2H, POLR2L, POLR3B, POLR3D, POLR3G, PRIM1, PRIM2, RRM1, RRM2, TYMS, UCK1, UCK2
KEGG	Pyrimidine metabolism	experimental (any)	25	7.305	8.39e-6	CTPS1, DCTD, DCTPP1, ENTPD4, ENTPD6, NME4, NME6, POLD2, POLR1A, POLR1B, POLR2A, POLR2C, POLR2D, POLR2H, POLR2L, POLR3B, POLR3D, POLR3G, PRIM1, PRIM2, RRM1, RRM2, TYMS, UCK1, UCK2
NCI	Validated targets of C-MYC transcriptional activation	experimental (weak)	22	6.173	2.02e-5	BCAT1, BIRC5, CCNB1, CCND2, CDC25A, CDCA7, DDX18, E2F3, EIF4A1, EP300, GAPDH, HMGA1, HSP90AA1, HUWE1, LIN28B, MYC, PDCD10, PEG10, PFKM, PMAIP1, POLR3D, TAF9
NCI	Validated targets of C-MYC transcriptional activation	experimental (any)	22	6.193	2.15e-5	BCAT1, BIRC5, CCNB1, CCND2, CDC25A, CDCA7, DDX18, E2F3, EIF4A1, EP300, GAPDH, HMGA1, HSP90AA1, HUWE1, LIN28B, MYC, PDCD10, PEG10, PFKM, PMAIP1, POLR3D, TAF9

Figure 1. (**A**) Result representation of the human miRNA centric view, restricted to miRNAs containing ‘let’. (**B**) Target gene results for hsa-let-7b-5p, the first miRNA from the results presented in Figure 1A. (**C**) Target pathway results for hsa-let-7b-5p, the first miRNA from the results presented in Figure 1A.



Number of significant miRNAs per pathway

In this table the number of significant miRNAs per pathway are depicted for the different evidence sets.

Show 25 entries Search:

Excel CSV Column visibility

Database	Pathway	Significant miRNAs (predicted - intersection)	Significant miRNAs (predicted - union)	Significant miRNAs (weak experimental evidence)	Significant miRNAs (strong experimental evidence)	Significant miRNAs (any experimental evidence)
KEGG	MicroRNAs in cancer	55	1431	52	48	117
KEGG	Pathways in cancer	30	1466	29	36	77
WikiPathways	Signaling Pathways in Glioblastoma	24	1106	24	26	76
WikiPathways	DNA Damage Response only ATM dependent	25	712	20	13	75

Pathways in cancer

miRNAs that are significantly enriched for this pathway

In this table miRNAs are depicted that have significantly more targets in this pathway than expected by chance.

Show 10 entries Search: experimental

Excel CSV Column visibility

miRNA	Evidence	Hits	Expected hits	P-value	Targets
hsa-miR-29b-3p	experimental (any)	30	4.638	3.00e-14	AKT2, BCL2, CASP8, CDC42, CDK6, COL4A1, COL4A2, COL4A5, COL4A6, FOS, GSK3B, ITGA6, ITGB1, LAMC2, MDM2, MMP2, MMP9, PDGFA, PDGFB, PDGFRA, PDGFRB, PIK3CG, PIK3R1, PPARD, PTEN, TGFB1, TGFB2, TGFB3, VEGFA, VHL
hsa-miR-29a-3p	experimental (any)	26	4.491	1.32e-10	ABL1, AKT2, BCL2, CASP8, CCND1, CDC42, CDK2, CDK4, CDK6, COL4A1, COL4A2, CRKL, FOS, IGF1, ITGA6, LAMC2, MDM2, MMP2, PDGFRB, PIK3R1, PTEN, RET, TGFB3, TRAF4, VEGFA, VHL
hsa-miR-145-5p	experimental (any)	20	4.302	8.49e-7	BRAF, CDK4, CDK6, E2F3, EGFR, EPAS1, ETS1, FZD6, HDAC2, IGF1R, MDM2, MMP1, MYC, NRAS, SMAD3, SMAD4, STAT1, TGFBR2, TPM3, VEGFA
hsa-miR-21-5p	experimental (any)	35	12.02	9.75e-7	AKT2, APC, APPL1, BCL2, CDK6, COL4A1, E2F1, E2F2, E2F3, EGFR, ERBB2, FAS, FASLG, FGF12, HIF1A, IGF1R, MMP2, MMP9, MSH2, MSH6, MYC, NFKB1, PIK3R1, PLD1, PTEN, PTK2, RB1, SKP2, STAT3, TGFB1, TGFB2, TGFBR2, VEGFA, VHL, WNT5A

Figure 2. (**A**) pathway centric view for human miRNAs and pathways. (**B**) Results of the pathway centric view for the second pathway from Figure 2A.

Especially for GO categories and miRNA sets, the hypergeometric distribution is known to be biased, as described by Bleazard *et al*. (23). In their paper, the authors argue that this bias might be caused by the many-to-many relationship between miRNA sets and associated targets. They even emphasize that the respective bias increases with the number of considered miRNAs in the test set. This means that for our considerations of single miRNAs the respective bias seems to have a less strong influence although it cannot be ruled out completely. Additionally, the representation of results as heat map helps to discover specific miRNA – pathway regulations.

Database implementation and functionality

Database implementation and updates. miRPathDB is designed as a document-oriented NoSQL database that provides a RESTful API and is connected to a user friendly web interface. The user interface is based on HTML5 and JavaEE technology using the Thymeleaf template engine, JQuery and AJAX. The database information is visualized using the DataTables plug-in for JQuery and the Highcharts JavaScript library.

The database is implemented in a manner that semi-automated update routines can be used to incorporate new results in regular intervals. Each new version of miRPathDB will get a new version number (currently 1.0), all pathway resources will be updated and all enrichment results will be recomputed. The update routines currently require 900 CPU hours of computing time. All compute intensive tasks are performed using the GeneTrail2 C++ library (https://github.com/unisb-bioinf/genetrail2) and GNU Parallel (27).

Database functionality

With the database we want to enable researchers to identify which miRNAs target a specific pathway, which pathways are regulated by a specific miRNA and how specific these

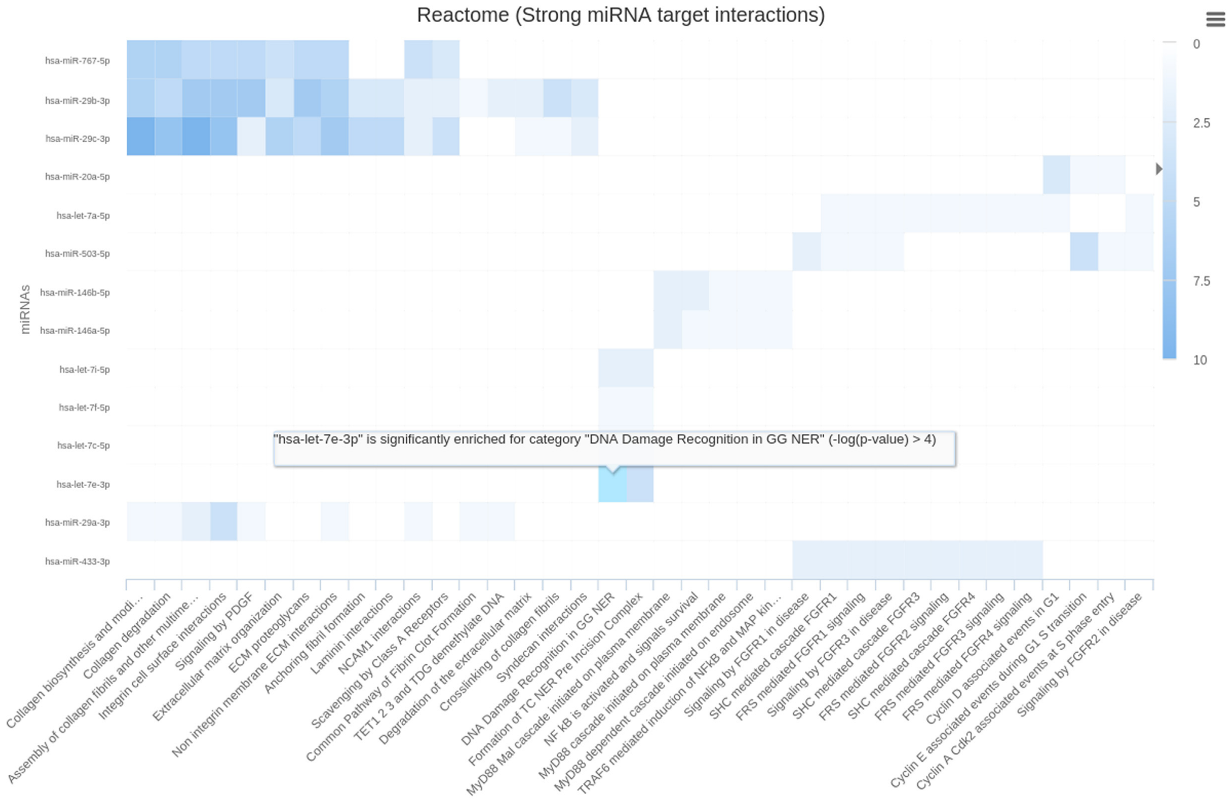


Figure 3. Interactive map for strong validated human target pathways in Reactome.

interactions are. miRPathDB contains data about molecular pathways and biological processes that are significantly more targeted by certain miRNAs than expected by chance. For each miRNA, we extracted five target sets (experimental evidence, strong experimental evidence, weak experimental evidence, intersection of predicted target data sets, union of predicted target data set) and pre-computed enrichment analyses for the categories mentioned above.

Following the goal described above we generated several representations of our database: a miRNA centric representation, a pathway centric representation and a detailed representation for each miRNA and pathway.

In the miRNA centric representation, all miRNAs are listed along with the number of significantly targeted categories with respect to the five target sets per miRNA. The user can sort miRNAs in alphabetic order or according to the number of targeted categories. Also, the results can be filtered by, e.g. typing 'let-7' in the search field, effectively, making it possible to inspect only results of the let-7 family. A typical result of the miRNA centric view is shown in Figure 1A. From here, users can select a miRNA of interest and obtain detailed information about its targets and regulated pathways. First, the target genes for this miRNA are listed and for each target the evidence(s) of this interaction is shown. The respective result is presented in Figure 1B. Per default, five entries are shown but the lists can be expanded to show, e.g. 50, 100, 250 or all target genes. Again, a gene name can be queried in the search field to see whether this gene is contained in the target gene list of the current miRNA. In addition to the target genes, the target pathways are listed. The basic set-up is similar to the target gene representation, however, in addition expected and actual number of target genes on the pathway are included, the respective adjusted significance value as well as the set of all genes that are targeted by this miRNA on that pathway. The representation of target pathways for hsa-let-7b-5p is presented in Figure 1C.

The pathway centric representation lists all pathways that are significantly enriched for at least one miRNA and one of the target sets. It follows generally the same scheme as the miRNA centric one: per pathway the number of miRNAs significantly targeting this pathway dependent on the different target set categories is listed. Again, the representation can be restricted to pathways containing a certain name, in the example in Figure 2A, only pathways with 'cancer' are listed. By selecting one pathway from the list the details are presented for this pathway. miRNAs with enriched number of target genes on that pathway are listed with the expected and actual number of target genes and the adjusted *P*-value followed by target gene names. An example is presented in Figure 2B. The detailed representation can be directly accessed for certain miRNAs or pathways by the search button in the upper right corner of the miRPathDB web page.

In order to complement this functionality, we also provide a graphic visualization of miRNA-pathway interactions as interactive heat maps that provide a comprehensive overview of pathways targeted by the different (single) miRNAs.

For each category and the five different target sets significance values are presented for miRNAs and the targeted pathways. The significance values are color-coded on a logarithmic scale. On mouse over, the *P*-value for the miRNA in the row and the pathway in that column is highlighted. An example for human target pathways of miRNAs with strong evidence target genes from Reactome is available in Figure 3. Using this representation, researchers can immediately see whether a miRNA is targeting only few pathways and is rather specific or whether a miRNA is targeting almost all categories.

Since miRNAs and their targets are conserved between organisms, we implemented the functionality described above for *Mus musculus* and *Homo sapiens*. Users can switch between the organisms by clicking on the respective organism logo located besides the search function in the upper right corner of the miRPathDB home page. Thereby, the degree of conservation of pathways between mouse and human can be assessed.

Download of results and data availability

In each of the miRNA and pathway centric representations the user can select the columns of interest and selectively hide information. The results can then be downloaded in common formats, including flat files (comma separated) that can be used as input for other tools or Excel lists. Beyond that, we also offer a download of the complete result tables from the miRPathDB homepage.

CONCLUSION

The increasing number of human miRNAs, availability of experimentally validated targets and updates in pathway resources lead to an altered picture of miRNAs targeting pathways. The complexity of the analyses calls for a concise and easy to use data repository storing the most recent interactions between miRNAs and target pathways. In the present study, we systematically analyzed target gene sets of miRNAs as well as the regulatory influence of miRNAs on pathways. With the miRNA Pathway Dictionary Database (miRPathDB), which is freely accessible at https://mpd.bioinf.uni-sb.de/, we provide a comprehensive collection of single miRNAs that regulate pathways, gene ontologies and other categories, hence complementing the hitherto available miRNA target enrichment programs, tailored for miRNA sets.

FUNDING

Funding for open access charge: Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Ambros,V. (2001) microRNAs: Tiny regulators with great potential. *Cell*, **107**, 823–826.
2. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111.
3. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
4. Akhtar,M.M., Micolucci,L., Islam,M.S., Olivieri,F. and Procopio,A.D. (2016) Bioinformatic tools for microRNA dissection. *Nucleic Acids Res.*, **44**, 24–44.
5. Chiromatzo,A.O., Oliveira,T.Y., Pereira,G., Costa,A.Y., Montesco,C.A., Gras,D.E., Yosetake,F., Vilar,J.B., Cervato,M., Prado,P.R. *et al.* (2007) miRNApath: A database of miRNAs, target genes and metabolic pathways. *Genet. Mol. Res.*, **6**, 859–865.
6. Kowarsch,A., Preusse,M., Marr,C. and Theis,F.J. (2011) miTALOS: Analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA*, **17**, 809–819.
7. Vlachos,I.S., Kostoulas,N., Vergoulis,T., Georgakilas,G., Reczko,M., Maragkakis,M., Paraskevopoulou,M.D., Prionidis,K., Dalamagas,T. and Hatzigeorgiou,A.G. (2012) DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Res.*, **40**, W498–W504.
8. Vlachos,I.S., Zagganas,K., Paraskevopoulou,M.D., Georgakilas,G., Karagkouni,D., Vergoulis,T., Dalamagas,T. and Hatzigeorgiou,A.G. (2015) DIANA-miRPath v3.0: Deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.
9. Stockel,D., Kehl,T., Trampert,P., Schneider,L., Backes,C., Ludwig,N., Gerasch,A., Kaufmann,M., Gessler,M., Graf,N. *et al.* (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, **32**, 1502–1508.
10. Capece,V., Garcia Vizcaino,J.C., Vidal,R., Rahman,R.U., Pena Centeno,T., Shomroni,O., Suberviola,I., Fischer,A. and Bonn,S. (2015) Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics*, **31**, 2205–2207.
11. Backes,C., Meese,E., Lenhof,H.P. and Keller,A. (2010) A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.*, **38**, 4476–4486.
12. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
13. Backes,C., Keller,A., Kuentzer,J., Kneissl,B., Comtesse,N., Elnakady,Y.A., Muller,R., Meese,E. and Lenhof,H.P. (2007) GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–W192.
14. Das,N. (2012) MicroRNA Targets - How to predict? *Bioinformation*, **8**, 841–845.
15. Vlachos,I.S., Paraskevopoulou,M.D., Karagkouni,D., Georgakilas,G., Vergoulis,T., Kanellos,I., Anastasopoulos,I.L., Maniou,S., Karathanou,K., Kalfakakou,D. *et al.* (2015) DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.*, **43**, D153–D159.
16. Chou,C.H., Chang,N.W., Shrestha,S., Hsu,S.D., Lin,Y.L., Lee,W.H., Yang,C.D., Hong,H.C., Wei,T.Y., Tu,S.J. *et al.* (2016) miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.*, **44**, D239–D247.
17. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
18. Stein,L.D. (2004) Using the Reactome database. *Curr. Protoc. Bioinformatics*, doi:10.1002/0471250953.bi0807s7.
19. Nishimura,D. (2004) BioCarta. *Biotech Software Internet Report*, **2**, 117–120.
20. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
21. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
22. Jewison,T., Su,Y., Disfany,F.M., Liang,Y., Knox,C., Maciejewski,A., Poelzer,J., Huynh,J., Zhou,Y., Arndt,D. *et al.* (2014) SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Res.*, **42**, D478–D484.
23. Bleazard,T., Lamb,J.A. and Griffiths-Jones,S. (2015) Bias in microRNA functional enrichment analysis. *Bioinformatics*, **31**, 1592–1598.

24. Paraskevopoulou,M. D., Georgakilas,G., Kostoulas,N., Vlachos,I. S., Vergoulis,T., Reczko,M., Filippidis,C., Dalamagas,T. and Hatzigeorgiou,A.G. (2013). DIANA-microT web server v5. 0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, **41**, W169–W173.
25. Wong,N. and Wang,X. (2014). miRDB: An online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.*, **43**, D146–D152.
26. Agarwal,V., Bell,G. W., Nam,J. W. and Bartel,D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
27. Tange, O. (2011) GNU Parallel - The Command-Line Power Tool. *login: The USENIX Magazine*, **2011**, 42–47.
28. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004). The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

3.20 *miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database*

D142–D147 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 6 November 2019*
doi: 10.1093/nar/gkz1022

miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database

Tim Kehl[1,†], Fabian Kern[2,†], Christina Backes[2], Tobias Fehlmann[2], Daniel Stöckel[1,3], Eckart Meese[4], Hans-Peter Lenhof[1] and Andreas Keller[2,5,6,*]

[1]Chair for Bioinformatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany, [2]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [3]EMD Digital, Merck KGaA, Darmstadt, Germany, [4]Department of Human Genetics, Saarland University, 66421 Homburg, Germany, [5]School of Medicine Office, Stanford University, Stanford, CA, USA and [6]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA

Received September 15, 2019; Revised October 17, 2019; Editorial Decision October 17, 2019; Accepted October 19, 2019

ABSTRACT

Since the initial release of *miRPathDB*, tremendous progress has been made in the field of microRNA (miRNA) research. New miRNA reference databases have emerged, a vast amount of new miRNA candidates has been discovered and the number of experimentally validated target genes has increased considerably. Hence, the demand for a major upgrade of *miRPathDB*, including extended analysis functionality and intuitive visualizations of query results has emerged. Here, we present the novel release 2.0 of the miRNA Pathway Dictionary Database (*miRPathDB*) that is freely accessible at https://mpd.bioinf.uni-sb.de/. *miRPathDB* 2.0 comes with a tenfold increase of pre-processed data. In total, the updated database provides putative associations between 27 452 (candidate) miRNAs, 28 352 targets and 16 833 pathways for *Homo sapiens*, as well as interactions of 1978 miRNAs, 24 898 targets and 6511 functional categories for *Mus musculus*. Additionally, we analyzed publications citing *miRPathDB* to identify common use-cases and further extensions. Based on this evaluation, we added new functionality for interactive visualizations and down-stream analyses of bulk queries. In summary, the updated version of *miRPathDB*, with its new custom-tailored features, is one of the most comprehensive and advanced resources for miRNAs and their target pathways.

INTRODUCTION

Understanding the mechanisms of gene regulation is one of the major challenges in molecular biology and bioinformatics. In order to get the big picture, diverse sub-fields emerged to study the underlying principles of transcriptional, post-transcriptional, translational and post-translational levels of gene regulation. Short, conserved and non-coding RNA families, so-called microRNAs (miRNAs), were shown to orchestrate major pathways in a post-transcriptional manner by targeting 3′ untranslated regions (UTRs) of mRNAs in mammals and plants (1,2). While early studies focused on the validation of human microRNAs and those found in important model organisms such as mouse and rat, the focus has been broadly expanded to characterize miRNAs in a larger set of metazoan species (3). To this end, several reference databases such as miRBase, miRCarta and miRGeneDB and different nomenclatures were established (4–6). Since the number of miRNAs discovered is steadily rising (7), a remarkable amount of studies already validated microRNA target genes and their function in a multitude of cell-types, tissues and disease phenotypes (8,9). These global research efforts have led to an accumulation of novel data. To scale up with these developments and to gain deeper insights into miRNA functionality, robust statistical methods and curated databases are in great demand, especially to integrate all the important findings from miRNA discovery, target validation and target gene function (10,11).

One of the key questions of functional miRNA analysis is which pathways or cellular functions are regulated by a given miRNA (miRNA-centric view), or conversely, which miRNAs regulate a given gene set or pathway (pathway-centric view) (12,13). To solve these problems, several tools and databases have been proposed so far. From a miRNA-centric view, the miRTar database, which links individual miRNAs to metabolic pathways (14) and miRSystem, providing pre-computed enrichments of target genes in pathways (15), should be noted. Moreover, pure enrichment-based tools like miEAA, the bioconductor package miRNApath, or BUFET that are based on many-to-many rela-

[*]To whom correspondence should be addressed: Tel: +49 681 302 68611; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

tionships can process lists of miRNA identifiers to compute pathway associations (16–18). More specialized miRNA-centric tools include miRNet (19), which is a networks-based approach and miTALOS v2 (20) that annotates miRNA functions in a tissue-specific manner. PolymiRTS (21) is a pathway-centric database that maps SNPs in target sites to gene categories and phenotypes, i.e. disease traits.

Only a minor fraction of the tools and databases support both miRNA- and pathway-centric applications. These include the online database miRNApath (22), the R package CORNA (23), DIANA-miRPath v3.0 (24) incorporating GO and KEGG enrichments derived from predicted and validated miRNA-target interactions, and finally *miRPathDB* v1 (25), which in turn is based on our very first dictionary on miRNAs and target pathways (26).

After the initial release of *miRPathDB*, miRNA research has made notable progress. Novel miRNAs have been discovered, the number of experimentally validated target genes has increased tremendously. Most importantly, new reference databases emerged that either catalog validated miRNAs with high confidence (6), or that contain thousands of novel miRNA candidates (5). Additionally, we evaluated publications, citing our database, to identify common application scenarios, new visualizations and useful downstream applications (27–29). An overview of these publications can be found in Supplementary Table S1.

The new version of *miRPathDB*, provides access to target genes and regulated pathways not only for miRNAs from miRBase (Version 22.1), but also from miRCarta (Version 1.1). This increases the provided information by more than a factor of ten compared to the original version. Second, our database now also provides similarity information for all miRNAs based on their sequence, genomic position, target genes and target pathways. This information not only allows to query miRNAs with similar properties and to cluster miRNAs based on their similarity, but also to assess the regulatory potential of new candidate miRNAs. On top of the new data compilation, *miRPathDB* provides several interactive tools for user-specific analyses. From a miRNA perspective, we developed an appealing miRNA-to-pathway heatmap visualization that intuitively shows which pathways are regulated by a given set of miRNAs. To serve the pathway-centric use-case as well, we have formulated and implemented an Integer Linear Program (ILP) to automatically extract a set of miRNAs whose targetome covers a user-provided pathway or set of genes. Taken together, the new version of *miRPathDB* is a comprehensive resource to study the function of miRNAs in human and mouse.

MATERIALS AND METHODS

Our database integrates information of miRNAs, miRNA–target interactions (MTIs), and signaling pathways from several third-party resources. In the following sections, we describe the respective data sources and all processing steps performed to create the underlying data collection. Additionally, we describe the methodology of new downstream analysis features.

miRNA resources

The database stores information on all human and mouse miRNAs from miRBase (Version 22.1) and from miRCarta (Version 1.1), including miRNA candidates. Validated MTIs were acquired from miRTarBase (Version 7) (30) and pre-processed to create two subsets for each miRNA: all MTIs independent of their type of experimental evidence and only those with a strong level of evidence. On top of this, we predicted target genes for each miRNA sequence using TargetScan (Version 7.1) (31) and MiRanda (Version 3.3a) (32). Based on the prediction output, we also created two further list of MTIs: the intersection and the union of all predictions, which is a common strategy to account for putative sources of bias from target prediction tools and to balance sensitivity versus specificity (25,33). As 3′ UTR input target set for the two algorithms, we used the curated annotations from *targetscan.org* for both human and mouse runs. Each program was executed using its default set of parameters.

Pathway databases and enrichment analysis

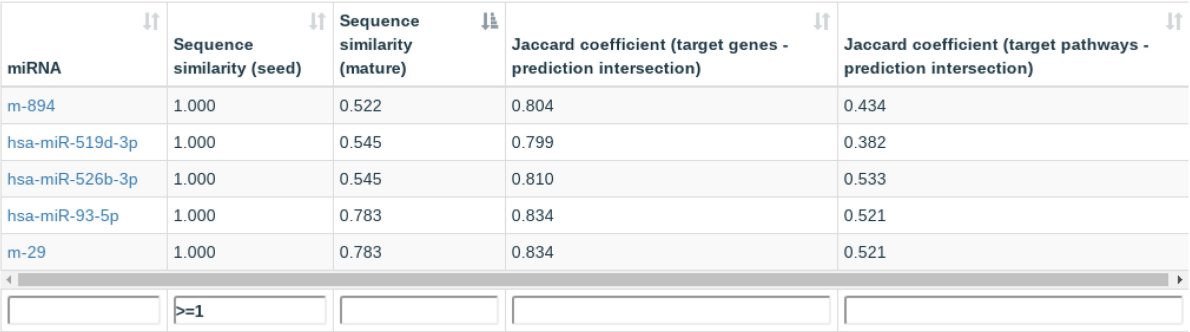
In order to determine whether a specific miRNA is associated with a particular biological process or signaling pathway, we used the enrichment analysis functionality of the GeneTrail2 C++ library (34). To this end, we analyzed functional categories from the Gene Ontology (35), as well as signaling pathways from KEGG (36), Reactome (37) and WikiPathways (38). For each pair of miRNA and functional category, we applied a hypergeometric test to check if the pathway contains significantly more target genes than expected by chance. Resulting p-values were FDR-adjusted (39) and a significance level of 0.05 was selected.

miRNA similarities

We also calculated similarities between all miRNAs and miRNA candidates based on their seed sequence, mature sequence, target genes and target pathways. For the string comparison, we calculated the Hamming distance between the sequences of all miRNA pairs, once using the full mature sequences and once the 7-nt substrings starting at position 2 from the 5′ end of the mature sequence. Given the hamming distance H_d between two sequences of length l, we defined the pairwise sequence similarity as $1 - (\frac{H_d}{l})$. The similarity of two sets containing either target genes or pathways was calculated using the Jaccard coefficient. Moreover, we compared miRNAs according to the positions of their genomic loci by computing the minimal distance between miRNAs annotated to the same chromosome.

Customized pathway heatmaps

The custom heatmap depicts which pathways are regulated by a user defined set of miRNAs. To create a heatmap, we first select all pathways that are significantly enriched for the targets of at least one of the specified miRNAs. The obtained p-values are used to construct a matrix that contains the $-\log_{10}$-transformed and discretized P-values for



miRNA	Sequence similarity (seed)	Sequence similarity (mature)	Jaccard coefficient (target genes - prediction intersection)	Jaccard coefficient (target pathways - prediction intersection)
m-894	1.000	0.522	0.804	0.434
hsa-miR-519d-3p	1.000	0.545	0.799	0.382
hsa-miR-526b-3p	1.000	0.545	0.810	0.533
hsa-miR-93-5p	1.000	0.783	0.834	0.521
m-29	1.000	0.783	0.834	0.521
	>=1			

Figure 1. Example of the new pairwise miRNA similarity table. The figure shows the pre-computed similarities for hsa-miR-106a-5p sorted by sequence similarity (mature) in increasing order. Furthermore, the table is filtered to show only miRNAs and miRNA candidates having 100% seed similarity. The Jaccard index provides additional information about the functional similarity of each miRNA and miR-106a-5p for predicted targets and target pathways.

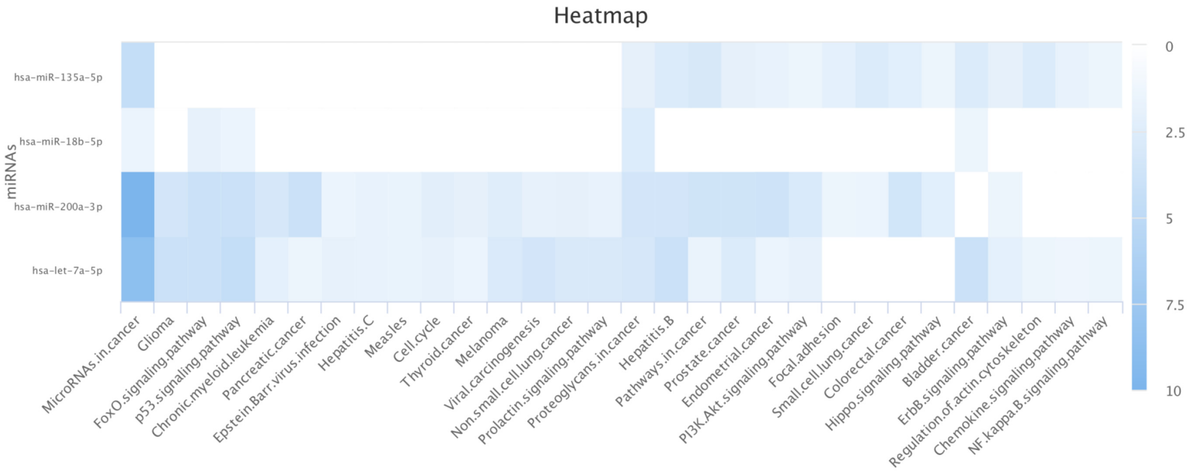


Figure 2. Example of the custom heatmap visualization. The figure depicts the enrichment results of hsa-miR-18b-5p, hsa-miR-135a-5p, hsa-let-7a-5p and hsa-miR-200a-3p for the categories of the KEGG database and strongly experimentally validated MTIs. Rows represent the enrichment results for the targets of the four miRNAs. Columns represent all KEGG pathways that are significant for the different miRNAs. For demonstration purposes, the heatmap was filtered to only show pathways with at least two associated miRNAs. The color of individual fields represent the $-\log_{10}$-transformed P-value of the respective enrichment results. Darker colors indicate more significant associations between miRNA and target pathway.

the set of miRNAs and all enriched pathways. Finally, similar miRNAs and pathways are clustered together by applying an hierarchical approach (Ward's method with Euclidian distance) to both rows and columns of the matrix. The clustered matrix is subsequently displayed as an interactive heatmap, implemented using the Highcharts JavaScript library.

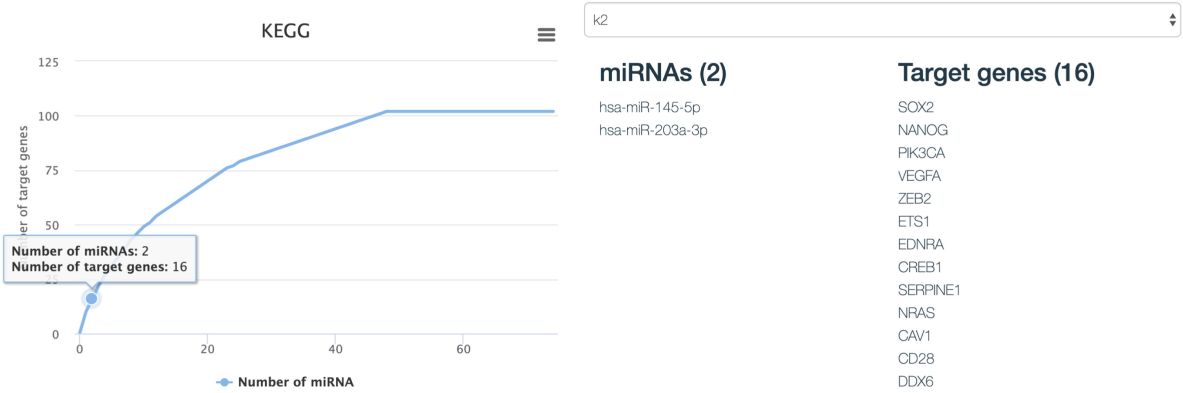
Maximum targetome coverage analysis

A noteworthy issue in functional miRNA research is to find a small number of miRNAs that are sufficient to regulate a given gene set, e.g. a particular signaling cascade or pathway. To solve this problem, we first search for the 'best' miRNA ($k = 1$) that regulates the maximal number of genes of the given target set. Next, we increase the considered number of miRNAs step-by-step ($k := k + 1$) until all target genes are covered or a predefined $k_{\max}$ is reached. For each k, we report an optimal set of miRNAs and the regulated target genes.

The problem to find the optimal set of miRNAs for one particular k is closely related to the maximum coverage problem, which can be solved using Integer Linear Programming (ILP). A formal definition of this problem can be found in the online documentation and Supplement S2. The ILP was implemented in C++ using the CPLEX optimization framework. Finally, results of an analysis are visualized by an interactive plot using the Highcharts JavaScript library.

OVERVIEW OF MIRPATHDB 2.0

miRPathDB stores information on (candidate) miRNAs, their target genes and their target pathways. To access this information, our database offers users two distinct representations: a miRNA-centric and a pathway-centric view. An overview table and a detailed description of each



miRNAs (2)	Target genes (16)
hsa-miR-145-5p	SOX2
hsa-miR-203a-3p	NANOG
	PIK3CA
	VEGFA
	ZEB2
	ETS1
	EDNRA
	CREB1
	SERPINE1
	NRAS
	CAV1
	CD28
	DDX6

Figure 3. Example of the interactive visualization for a user-specific maximum-coverage analysis. The curve on the left indicates how many of the specified target genes can be targeted by an increasing number of miRNAs. Here the x-axis shows the increasing number of miRNAs and the Y-axis the number of covered target genes. Users are able to click on every point of the curve to inspect the corresponding miRNAs and targeted genes. An example for $k = 2$ is depicted on the right-hand side.

miRNA or pathway are available. The representations can either be accessed through the overview tables or a query in the quick-search bar. A general description of these representations has already been presented in the original manuscript (26). Hence, we describe the extensive changes of the miRNA-centric view, the interactive analysis tools, and the new export functionality in the following sections.

NEW MIRNA-CENTRIC VIEW

Here, we explain the different levels of information *miRPathDB* offers for each miRNA or miRNA candidate.

General information

On the top of each miRNA page, we provide general information about the respective miRNA: the precursor mapping, the sequence of the mature miRNA, seed and corresponding parent stem loops, and all annotated genomic loci. Additionally, specific links to external reference database (miRBase and miRCarta) entries for all miRNAs and for corresponding precursors and family assignments are deposited. On top of this, each miRNA entry is linked to other third-party databases, like the TissueAtlas (40) or miRTargetLink (41), not only to improve the usability, but also to complement the features of *miRPathDB* with other essential tools for miRNA analysis.

miRNA-target interactions (MTIs)

Below the general information section, the website renders a responsive, sortable, and fully searchable table containing all target genes of an examined miRNA. For each gene, we also highlight in which of the four evidence sets it is contained. Table rows can be filtered using the text boxes below each column. Users can export both filtered and unfiltered tables in different file formats (CSV, Excel and PDF).

Targeted pathways

One major focus of our database is to provide information on associations between miRNAs and their putative target pathways. Likewise to the table for target genes, the pathways are shown in another fully responsive table. It contains, for the different evidence sets, all pathways that are significantly enriched with targets of the examined miRNAs. For each pathway, the number of contained target genes, the number of target genes that are expected by chance, and a FDR-adjusted *P*-value are listed. Since users might be interested in a specific subset of results, the table can be filtered with respect to all fields. For example, users can select significant pathways for a certain MTI evidence level, or only pathways that contain a specific gene of interest. Each pathway cell is linked to the corresponding external database entry, which often displays additional information like a description of the pathway or the underlying gene network.

miRNA similarities

At the bottom of each miRNA page, a novel table containing similarity information of the selected miRNA with respect to all other miRNAs from the same organism, including miRNA candidates, is displayed (Figure 1). The table lists the seed and full sequence similarities, the chromosomal distance, in case miRNAs are annotated on the same chromosome and eight Jaccard coefficients, measuring the similarity of target genes and target pathways for the different evidence sets of MTIs. Analogously to information about target genes and pathways, this table can be filtered, searched, sorted, resized, and exported for further usage.

NEW INTERACTIVE DATABASE FUNCTIONALITY

In addition to a new data compilation, *miRPathDB* features several new interactive tools for advanced user-specific database queries and analyses.

Custom pathway heatmaps

A common question in miRNA research is, whether targets of deregulated miRNAs are similarly enriched in certain biological processes or are associated with distinct molecular functions (27,28). In order to help users to tackle this question, we developed an interactive heatmap visualization. To create this plot, a user needs to specify a list of miRNAs as well as the evidence level for the MTIs. *miRPathDB* automatically selects all functional categories that are significantly enriched for the targets of at least one of the specified miRNAs. Results are represented as a heatmap, where each row depicts enrichment results for the respective functional categories. The color of individual entries corresponds to the p-value of the associated enrichment result. Darker colors indicate more significant enrichments of miRNA target genes in the corresponding biological processes. On top of this, users may specify the resolution of the resulting heatmap and download the image in different file formats (PNG, JPEG, PDF and SVG). An example heatmap is shown in Figure 2. Our customized heatmap feature provides a rapid overview of molecular functions and signaling pathways that are potentially regulated by a specific miRNA set. This analysis might even be helpful to assess possible downstream effects of deregulated miRNAs in high-throughput studies.

Maximum targetome coverage analysis

While the previous feature allows downstream analysis from a miRNA-centric view, by mapping a given miRNA set to enriched target pathways, *miRPathDB* also provides functionality for the reverse direction, i.e. given a set of target genes G, find a minimal set of miRNAs that target all genes in G. To this end, we provide a tool that iteratively computes k miRNAs (for all $k \in \{1, 2, ..., k_{\max}\}$) with a maximal number of targets in G (see Materials and Methods). To start the maximum coverage analysis, a user must upload a list of genes, select the desired level of evidence that should be used to lookup the MTIs and set the largest $k = k_{\max}$ where the algorithm should stop. The results of such an analysis are displayed in an interactive line-graph that plots k against the number of covered target genes (Figure 3, left). For each k, a node is inserted in the graph that can be selected. Upon selection of a node, the website displays an optimal set of miRNAs of the corresponding size k along with the list of overlapping target genes (Figure 3, right).

DATA EXPORT

Most of the views in *miRPathDB* offer dedicated export functionality. All tables in the miRNA-centric and the pathway-centric view can be filtered and downloaded in different formats (CSV, Excel and PDF). Additionally, we host downloads for all processing steps of the enrichment analyses. Users are able to acquire the unprocessed enrichment results, i.e. a table containing detailed information for each functional category. Furthermore, a table containing all pairs of miRNA and pathways and their $-log_{10}$-transformed p-values is available. *miRPathDB* also supplies all functional categories in Gene Matrix Transposed (GMT) format (cf. Online documentation).

CONCLUSION

Recent advancements in miRNA research yielded huge numbers of novel miRNAs, miRNA candidates, and experimentally validated MTIs. This circumstance motivated a novel release of *miRPathDB*. Besides miRNAs and their targets, our database also provides information about associations between pathways and miRNAs. Beyond the tenfold increase of data, our database now offers powerful tools for the visualization and downstream analysis of database queries. In particular, users are able to search similar miRNAs, create interactive clustered heatmaps and to determine a minimal set of candidate regulators that are sufficient to target a specified gene list. In summary, *miRPathDB* 2.0 is the most comprehensive publicly available resource to assess the relationship between microRNAs, their targets and cellular functions for human and mouse.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Internal funds of Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Jonas,S. and Izaurralde,E. (2015) Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.*, **16**, 421–433.
2. Bartel,D.P. (2018) Metazoan microRNAs. *Cell*, **173**, 20–51.
3. Fehlmann,T., Backes,C., Pirritano,M., Laufer,T., Galata,V., Kern,F., Kahraman,M., Gasparoni,G., Ludwig,N., Lenhof,H.P. *et al.* (2019) The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Res.*, **47**, 4431–4441.
4. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
5. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.P., Meese,E. and Keller,A. (2018) MiRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
6. Fromm,B., Domanska,D., Høye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2019) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, doi:10.1093/nar/gkz885.
7. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grässer,F.A., Lenhof,H.P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, **47**, 3353–3364.
8. Rupaimoole,R. and Slack,F.J. (2017) MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.*, **16**, 203–221.
9. Karagkouni,D., Paraskevopoulou,M.D., Chatzopoulos,S., Vlachos,I.S., Tastsoglou,S., Kanellos,I., Papadimitriou,D., Kavakiotis,I., Maniou,S., Skoufos,G. *et al.* (2018) DIANA-TarBase v8: A decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.*, **46**, D239–D245.
10. Liu,B., Li,J. and Cairns,M.J. (2012) Identifying miRNAs, targets and functions. *Brief. Bioinformatics*, **15**, 1–19.
11. Fehlmann,T., Laufer,T., Backes,C., Kahramann,M., Alles,J., Fischer,U., Minet,M., Ludwig,N., Kern,F., Kehl,T. *et al.* (2019) Large-scale validation of miRNAs by disease association, evolutionary conservation and pathway activity. *RNA Biol.*, **16**, 93–103.
12. Davis,J.A., Saunders,S.J., Mann,M. and Backofen,R. (2017) Combinatorial ensemble miRNA target prediction of co-regulation

networks with non-prediction data. *Nucleic Acids Res.*, **45**, 8745–8757.

13. Sticht,C., De La Torre,C., Parveen,A. and Gretz,N. (2018) miRWalk: an online resource for prediction of microRNA binding sites. *PLoS ONE*, **13**, 1–6.
14. Hsu,J.B., Chiu,C.M., Hsu,S.D., Huang,W.Y., Chien,C.H., Lee,T.Y. and Huang,H.D. (2011) MiRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics*, **12**. 300.
15. Lu,T.P., Lee,C.Y., Tsai,M.H., Chiu,Y.C., Hsiao,C.K., Lai,L.C. and Chuang,E.Y. (2012) miRSystem: An Integrated System for Characterizing Enriched Functions and Pathways of MicroRNA Targets. *PLoS ONE*, **7**, 1–10.
16. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) MiEAA: MicroRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
17. Cogswell,J.P., Ward,J., Taylor,I.A., Waters,M., Shi,Y., Cannon,B., Kelnar,K., Kemppainen,J., Brown,D., Chen,C. *et al.* (2008) Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways. *J. Alzheimer's Dis.: JAD*, **14**, 27–41.
18. Zagganas,K., Vergoulis,T., Paraskevopoulou,M.D., Vlachos,I.S., Skiadopoulos,S. and Dalamagas,T. (2017) BUFET: boosting the unbiased miRNA functional enrichment analysis using bitsets. *BMC Bioinformatics*, **18**, 399.
19. Fan,Y., Siklenka,K., Arora,S.K., Ribeiro,P., Kimmins,S. and Xia,J. (2016) miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res.*, **44**, W135–W141.
20. Preusse,M., Theis,F.J. and Mueller,N.S. (2016) miTALOS v2: analyzing tissue specific microRNA function. *PLoS ONE*, **11**, 1–15.
21. Bhattacharya,A., Ziebarth,J.D. and Cui,Y. (2013) PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res.*, **42**, D86–D91.
22. Chiromatzo,A.O., Oliveira,T.Y.K., Pereira,G., Costa,A.Y., Montesco,C.A.E., Gras,D.E., Yosetake,F., Vilar,J.B., Cervato,M., Prado,P.R.R. *et al.* (2007) miRNApath: a database of miRNAs, target genes and metabolic pathways. *Genet. Mol. Res.: GMR*, **6**, 859–865.
23. Wu,X. and Watson,M. (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics*, **25**, 832–833.
24. Vlachos,I.S., Zagganas,K., Paraskevopoulou,M.D., Georgakilas,G., Karagkouni,D., Vergoulis,T., Dalamagas,T. and Hatzigeorgiou,A.G. (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.
25. Backes,C., Kehl,T., Stöckel,D., Fehlmann,T., Schneider,L., Meese,E., Lenhof,H.P. and Keller,A. (2017) MiRPathDB: A new dictionary on microRNAs and target pathways. *Nucleic Acids Res.*, **45**, D90–D96.
26. Backes,C., Meese,E., Lenhof,H.P. and Keller,A. (2010) A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.*, **38**, 4476–4486.
27. Denham,J., Gray,A.J., Scott-Hamilton,J., Hagstrom,A.D. and Murphy,A.J. (2018) Small non-coding RNAs are altered by short-term sprint interval training in men. *Physiol. Rep.*, **6**, e13653.
28. Ragni,E., De Luca,P., Perucca Orfei,C., Colombini,A., Viganò,M., Lugano,G., Bollati,V. and de Girolamo,L. (2019) Insights into inflammatory Priming of adipose-derived mesenchymal stem cells: validation of extracellular vesicles-embedded miRNA reference genesas a crucial step for donor selection. *Cells*, **8**, E369.
29. Kehl,T., Backes,C., Kern,F., Fehlmann,T., Ludwig,N., Meese,E., Lenhof,H.P. and Keller,A. (2017) About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget*, **8**, 107167–107175.
30. Chou,C.H., Shrestha,S., Yang,C.D., Chang,N.W., Lin,Y.L., Liao,K.W., Huang,W.C., Sun,T.H., Tu,S.J., Lee,W.H. *et al.* (2018) MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **46**, D296–D302.
31. Agarwal,V., Bell,G.W., Nam,J.W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**. doi:10.7554/eLife.05005.
32. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
33. Bhattacharya,A. and Cui,Y. (2015) MiR2GO: comparative functional analysis for microRNAs. *Bioinformatics*, **31**, 2403–2405.
34. Stöckel,D., Kehl,T., Trampert,P., Schneider,L., Backes,C., Ludwig,N., Gerasch,A., Kaufmann,M., Gessler,M., Graf,N. *et al.* (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, **32**, 1502–1508.
35. The Gene Ontology Consortium (2018) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
36. Kanehisa,M., Sato,Y., Furumichi,M., Morishima,K. and Tanabe,M. (2018) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
37. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B. *et al.* (2017) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
38. Slenter,D.N., Kutmon,M., Hanspers,K., Riutta,A., Windsor,J., Nunes,N., Mélius,J., Cirillo,E., Coort,S.L., Digles,D. *et al.* (2017) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
39. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodological)*, **57**, 289–300.
40. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stähler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.
41. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) MiRTargetLink-miRNAs, genes and interaction networks. *Int. J.f Mol. Sci.*, **17**, 564.

3.21 *A review of databases predicting the effects of SNPs in miRNA genes or miRNA-binding sites*

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

This article is available under: https://doi.org/10.1093/bib/bbx155

3.22 *miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems*

Published online 6 May 2020 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* ***W521–W528***
doi: 10.1093/nar/gkaa309

miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems

Fabian Kern[1,†], Tobias Fehlmann[1,†], Jeffrey Solomon[1], Louisa Schwed[1], Nadja Grammes[1], Christina Backes[1], Kendall Van Keuren-Jensen[2], David Wesley Craig[3], Eckart Meese[4] and Andreas Keller[1,5,6,*]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA, [3]Institute of Translational Genomics, University of Southern California, Los Angeles, CA 90033, USA, [4]Department of Human Genetics, Saarland University, 66421 Homburg, Germany, [5]School of Medicine Office, Stanford University, Stanford, CA 94305, USA and [6]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94304, USA

Received March 05, 2020; Revised April 06, 2020; Editorial Decision April 17, 2020; Accepted April 22, 2020

ABSTRACT

Gene set enrichment analysis has become one of the most frequently used applications in molecular biology research. Originally developed for gene sets, the same statistical principles are now available for all omics types. In 2016, we published the miRNA enrichment analysis and annotation tool (miEAA) for human precursor and mature miRNAs. Here, we present miEAA 2.0, supporting miRNA input from ten frequently investigated organisms. To facilitate inclusion of miEAA in workflow systems, we implemented an Application Programming Interface (API). Users can perform miRNA set enrichment analysis using either the web-interface, a dedicated Python package, or custom remote clients. Moreover, the number of category sets was raised by an order of magnitude. We implemented novel categories like annotation confidence level or localisation in biological compartments. In combination with the miRBase miRNA-version and miRNA-to-precursor converters, miEAA supports research settings where older releases of miRBase are in use. The web server also offers novel comprehensive visualizations such as heatmaps and running sum curves with background distributions. We demonstrate the new features with case studies for human kidney cancer, a biomarker study on Parkinson's disease from the PPMI cohort, and a mouse model for breast cancer. The tool is freely accessible at: https://www.ccb.uni-saarland.de/mieaa2.

INTRODUCTION

Transcriptomics designates an indispensable set of techniques to study gene expression, often in a genome-wide manner, as the backbone of modern molecular biology and clinical research. The innumerable amount of classical bulk-sequencing datasets is further augmented by the recent advancements in high-resolution single-cell approaches. Since gene expression is constituted by many biological factors, experimental focus has been enlarged to include the regulatory non-coding transcriptome (ncRNAs), i.e. to RNA classes that regulate messenger RNAs (mRNAs) either directly or indirectly. Among these, microRNAs (miRNAs) are small non-coding RNAs, typically 18-25 nucleotides in length, loaded into proteins of the AGO-family to build RNA-induced silencing complexes (RISC) (1). Gene regulation through the RISC complex is facilitated by one or two mature ($-5p$; $-3p$) miRNA arms, arising from one or several transcribed precursors (2). Besides other modes of action, activated complexes target preferentially $3'$-untranslated regions of mRNAs to induce either catalytic cleavage or translation repression. Hence, profiling miRNA expression contributes to the understanding of gene regulation and potentially portrays cellular states. To date, numerous studies highlight their informative role in disease detection, sub-type classification, or progression, such as for cancer (3), neurodegenerative (4), or metabolic disorders (5) with a variety of bio-specimens (6).

Considering that several thousands of miRNAs have already been discovered, many novel miRNA candidates have been additionally proposed (7), while the total number of human miRNAs is estimated to be 2300 (8). Finding differences in expression for miRNAs is similar to mRNAs

[*]To whom correspondence should be addressed. Tel: +49 681 302 68611; Email: andreas.keller@ccb.uni-saarland.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

and therefore non-trivial. Differential gene expression studies often lead to dozens, hundreds, or even thousands of deregulated genes. Thus, large scale studies often make use of the functionality of gene set enrichment analysis (GSEA) (9). GSEA can further reduce large amounts of information towards a significant set of molecular functions, biological properties, or pathways of genes. In principle, a user inputs either a set or ordered list of genes and the tool runs the required statistical algorithms and provides background datasets to compare against.

Similar functionality was also implemented for other omics types, including proteomics, metagenomics or epigenomics. An in-depth review of gene set analysis methods for data other than mRNAs demonstrates the increasing interest and demand of the community in respective tools (10). We previously developed a statistical approach tailored for both miRNA precursor and mature miRNA input, the miRNA enrichment analysis and annotation tool (miEAA) (11). Here, we present an update of this tool that includes more categories, supports nine additional species, has new statistical functionality and offers a standardised Application Programming Interface (API) to facilitate the inclusion in modern data analysis workflows (12).

Given the growing interest in miRNAs, other tools with similar functionality to miEAA exist. The pioneering tool providing functionality for miRNA enrichment was TAM (13), which covers in it's latest version 2.0 (14) as many as 1238 human miRNA categories obtained from manual literature review of ∼9000 scientific manuscripts, along with new query and visualization features. In addition to the over- and under-representation analysis, users can compare the correlation of two miRNA lists under different disease conditions. Another important tool with similar functionality is miSEA (microRNA Set Enrichment Analysis) (15). It facilitates the selection of a large set of microRNA categories, including family classification, disease association, and genome coordinate. Furthermore, custom miRNA sets can be defined by the user. All kinds of enrichment tools rely on high quality sets of miRNA categories that were either obtained by curation of scientific literature or collected from specific databases. For instance, curated miRNA annotations can be obtained from miRBase (16) or miRCarta (17), miRNA–target interactions from miRTarBase (18), miRNA–pathway associations from miRPathDB (19), tissue-specific miRNAs from the human TissueAtlas (20), or miRNA-disease associations from HMDD (21) or MNDR (22), many of which were updated in the last two years. Further specialized annotations like miRNA and transcription factor interactions from TransmiR (23), miRNA sub-cellular localisations collected in RNALocate (24), or extra-cellular circulating miRNAs contained in miRandola (25) provide target categories for comprehensive enrichment analysis.

MATERIALS AND METHODS

In miEAA 2.0, we provide support for ten species whereas the first release of miEAA only supported *Homo sapiens*, 31 new category sets, and updates to our pre-existing datasets. To unify data preprocessing, we implemented an automated pipeline using Snakemake (26), Python 3.6, and the pandas (27) Python package facilitating data collection and filtering steps. For each species and their corresponding data sources our pipeline performs the same basic process, consisting of downloading the datasets, cleaning and updating the miRNA and precursor identifiers, transforming the results into a Gene Matrix Transposed (GMT) file, and creating background reference sets. Files were copied to the web server without further modification.

Data collection

Novel datasets were obtained to build our enrichment categories, consisting of Gene Ontology (28), miRTarBase 8.0 (18), KEGG (29), miRandola 2017 (25), miRPathDB 2.0 (19), TissueAtlas (20), MNDR v2.0 (22), NPInter 4.0 (30), RNALocate v2.0 (24), SM2miR (31), TAM 2.0 (14) and TransmiR v2.0 (23). Further annotations for cell-type and tissue specific expression of miRNAs and precursors were derived from three dedicated atlas publications (32,33) (10.1101/430561). Other pre-existing datasets have been updated, including HMDD v3.0 (21) and miRBase v22.1 (16). We retained the rest of our pre-existing datasets, namely miRWalk2.0 (34), published age and gender dependent miRNAs and distribution of miRNAs in immune cells (11). Most of the datasets contain miRNAs or precursors for *H. sapiens*. When available, we also utilise the data to derive categories representing the non-human organisms. Raw datasets were obtained either through a direct download or via an API. In particular, the QuickGO and KEGG datasets are compiled by querying corresponding REST APIs.

Category data preprocessing

First, data from QuickGO was mapped back to miRBase using RNAcentral (35). NCBI Gene was used in conjunction with miRTarBase to produce the indirect annotations. With the aid of the miRBaseConverter R package (36), miRNA and precursor names were translated to the latest version of miRBase. For KEGG Pathways and GO Annotations (direct and indirect through target genes from miRTarBase) we only keep miRNAs for which functional MTI support is available. In the MNDR diseases category set, we exclude HMDD data as it is precursor based, and MNDR is for mature miRNAs. To determine tissue-specific expression we computed the tissue specificity index (20) and applied a threshold filtering at 0.75.

Web server, statistics, and API implementation

The miEAA web server was built using a dockerized Django Web Framework v2.1, which exposes a web-API using the Django REST framework. The celery software was used as the job scheduler. Frontend libraries comprise Highcharts, dataTables, jquery, and Bootstrap. *P*-value correction methods were implemented using the R stats package. As gene set enrichment analysis (GSEA) implementation we provide an un-weighted variant of the algorithm. This implies the amount by which the running sum is changed in each step is constant, corresponding to a Kolmogorow–Smirnow test. This approach enables to compute the exact *P*-value without requiring permutations of either the case / control labels, or the miRNA lists (37). As an exception, the static GSEA running sum plots are computed by randomly permuting the test set 100 times and traversing the running sum for each random permutation. If the absolute maximal deviation from zero is positive, miEAA assumes an enrichment on top of the ordered list and results are shown in red colour to denote an enrichment. If the absolute maximal deviation from zero is negative, miEAA assumes an enrichment at the end of the ordered list and results are displayed in green color to denote an inverse enrichment, i.e. a depletion. Alongside our new API we provide a lightweight Python package, as well as a command line interface (CLI) tool, supporting Python 3.5 or higher. These are made freely available through the Python Package Index (pip) and through the *ccb-sb* conda channel. The already existing miRNA to precursor and miRBase converters were upgraded to miRBase v22.1. The former offers new output modes to simplify the review of ambiguous conversion results and proper down-stream usage.

Case studies

Raw and reads per million miRNA mapping (rpmmm) normalized miRBase v21 precursor counts and metadata of kidney renal clear cell carcinoma case and control samples were obtained from The Cancer Genome Atlas (TCGA). Since multiple sequencing results might be associated with the same sample ID in TCGA, we kept only one result file for each sample by preferring files from H over R over T analytes and selecting the aliquot with the highest plate number and / or lexicographical sorting order. Subsequently, miRNAs with fewer than 5 raw reads in less than 50% of either case or control samples were discarded from the analysis. All remaining miRNA counts were $\log_2$-scaled. Effect size was calculated using the implementation of Cohen's d from the R package effsize. Lists of precursor names, either selected by statistical significance or ordered by effect size, were converted from miRBase v21 to v22.1 using the online miRBase converter feature of miEAA. The list of all precursors from miRBase v21, converted to v22.1, were used as a reference set. The configured parameters included default precursor category sets without the *PubMed ID* and *TransMiR Tissues* sets, BH-FDR adjustment to a significance level of 0.05 with independently adjusted *P*-values per category set, and a minimum of 2 required hits per subcategory.

For the second case study, raw Agilent microarray data and sample metadata was downloaded from NCBI's GEO using accession ID GSE117000. Array parsing and probe signal processing was performed identically to the description in the first publication of miEAA (11). Subsequently, all counts were quantile-normalized and $\log_2$-transformed. All further down-stream analyses were performed analogous to the first case-study described above.

To provide a non-cancer case study we evaluated the performance of miEAA on a high-resolution dataset of small non-coding RNAs in whole blood (38). This dataset is freely available from the Parkinson's Progression Markers Initiative (PPMI) data portal. In summary, for 1600 individuals up to five blood samples from a time frame of over three years were acquired and sequenced for sncRNAs. We quantified all human miRBase v22 precursors from the 4340 sequencing samples. Raw counts were normalized to reads per million (rpm) and precursors were filtered analogously to the criteria defined for the TCGA case study. Next, we compared the miRNA precursor profiles of 2337 Parkinson's samples to 1538 age-matched controls. For this case study we also mapped back the precursors to miRBase v21 to perform a detailed comparison of enrichment results to TAM 2.0.

RESULTS

Overview on miEAA 2.0

In the following, changes and novelties introduced by the second major release of miEAA are described. Since all annotations of miRNAs to categories and databases are with respect to the miRNA reference database, miRBase, we converted the datasets to match its latest public version 22.1. This also affects the miRBase-version and miRNA-to-precursor converters, the former of which was designed to be fully backwards compatible. Moreover, both ORA and GSEA algorithms accept lists of either precursors or miRNAs, from *H. sapiens*, *Mus musculus*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Bos taurus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Gallus gallus* and *Sus scrofa*. In total, 134 525 categories from 16 published databases/resources are available to test against. A detailed breakdown of the counts by source and organism, on database and category set level, are available from Supplementary Table S1 and S2, respectively. For the precursor annotations, we curated family assignments, re-computed genomic clusters of miRNA genes, updated the chromosomal locations for human, and added all similar categories for other species. We also updated the category set representing PubMed IDs of manuscripts that contributed miRNA entries to miRBase. This feature has both, a biological and technical aspect. From the technical view, miRNAs could have been reported by the original paper due to experimental bias. In case a new input query is enriched for respective miRNAs it could be due to the same kind of bias. From a biological perspective, a study might have found miRNAs in the context of a disease. If such a manuscript is identified in a similar context in miEAA, additional evidence for the validity can be inferred. All species except *A. thaliana* are annotated with a new category listing high confidence precursors according to miRBase criteria. For human data, we transferred the disease annotations from HMDD to the new major release v3. We added associations from MNDR to allow disease comparisons against HMDD, and incorporated functional RNA interactions from NPInter. Lastly, novel categories such as the cellular localisation of miRNAs and regulatory interactions between miRNAs and transcription factors were incorporated from RNALocate and TransmiR, respectively. For the mature miRNAs, comparable changes apply as for the precursors in the cases of miRBase, MNDR, NPInter, and RNALocate-derived category sets. The gap between annotations of miRNA properties and their function is filled by categories on target genes taken from miRTarBase. Moreover, known miRNA to drug assocations are provided from SM2miR. To facilitate target-based enrichment of molecular pathways or biological function, we computed enrichments on target genes of miRNAs using Gene Ontology and KEGG. As an alternative for end-users, pre-computed significant enrichments of miRNAs associated with pathways provided by miRPathDB were made available for analysis. As the data from miRPathDB already involves a statistical pre-filtering, we implemented a new list of expert categories to highlight the underlying differences. Manually curated classifications from miRandola about known circular or extracellular miRNAs are also integrated. Finally, new annotations for cell-type and tissue-specific precursors and miRNAs have been integrated. Supposedly, the substantially enlarged number of categories might increase the average runtime of our algorithms, especially for the computationally intensive GSEA. Therefore, we profiled and improved our GeneTrail-based implementation to be three times faster, on average (39).

We raised the available number of statistical parameter settings as well. First, users can request unadjusted or adjusted P-values using six published techniques to account for multiple hypothesis testing on the same dataset. In addition to the classical Bonferroni and Benjamini–Hochberg False discovery rate (BH-FDR) procedures, the adjustments proposed by Benjamini-Yekuteli, Hochberg, Holm and Hommel can be selected. Moreover, the default behavior of miEAA to correct P-values database / category setwise was extended by a P-value pooling approach. In summary, the well-established alternatives for P-value correction can support highly customized research setups where alternate levels of stringency are required (40).

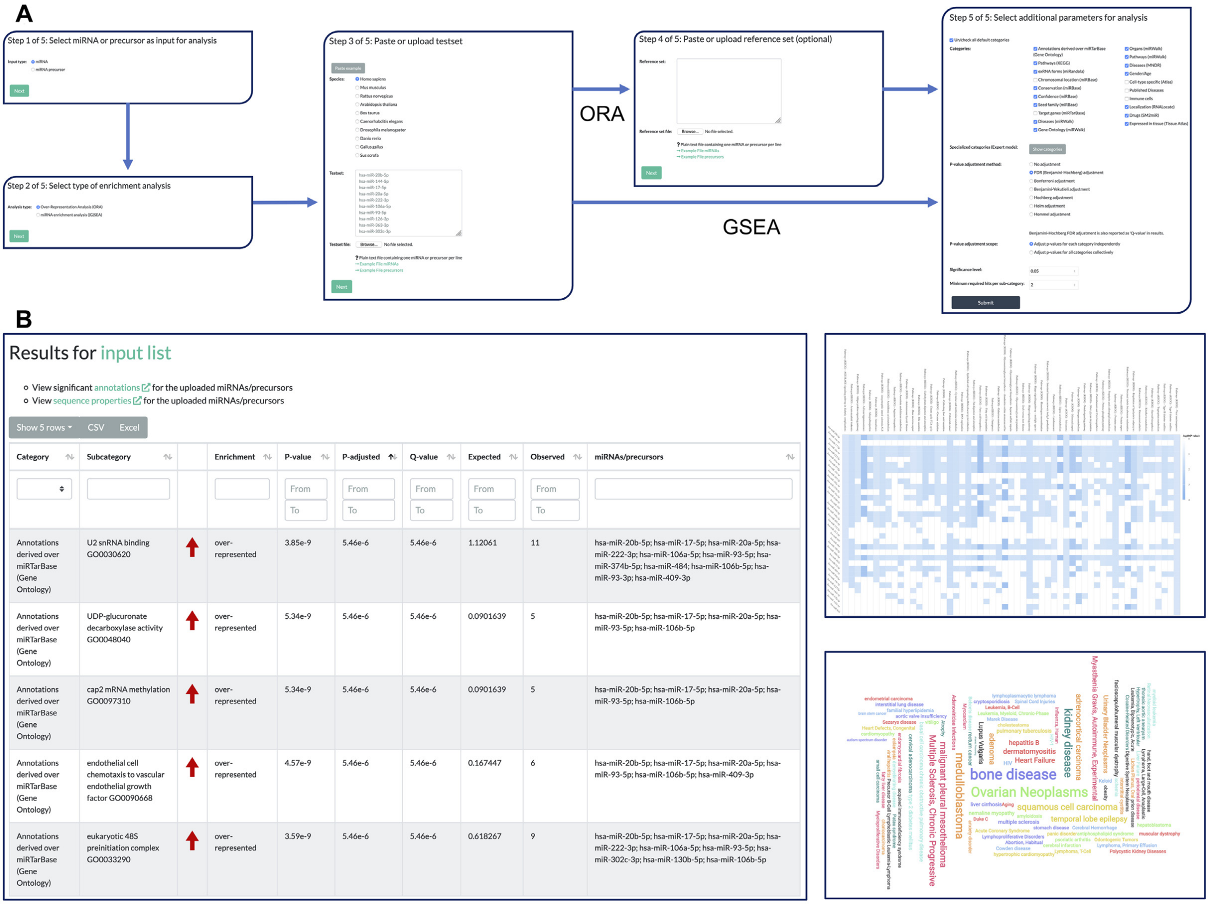
We also evaluated new visualization features for the output of enrichment analyses to provide a simple overview and to improve comprehension. As a result, we made existing graphs interactive and implemented enrichment graphs with simulated background distributions for GSEA as well as automatic word cloud and heatmap plots for all enrichment algorithms. Word clouds display the names of obtained categories while scaling the size of the terms relatively to the number of hits that occurred (on a linear or logarithmic scale) and allow one to qualitatively compare the categories. On top of that, category to miRNA heatmaps depict log-transformed P-values for the hits obtained. This feature permits to compare the similarity of enriched / depleted categories with respect to associated miRNAs or precursors in a simple fashion. The workflow of miEAA and example visualizations are displayed in Figure 1. Finally, we enhanced the general accessibility of miEAA through the implementation of a public API and a Python package, for which more details are provided below.

Case study 1: Human kidney renal clear cell carcinoma

As the first case-study of miEAA 2.0, we acquired 591 human miRNA-seq samples from the kidney renal clear cell carcinoma (KIRC) project of TCGA, which can be divided into 520 Primary tumor (PT) and 71 Solid tissue normal (STN) samples. Sample information can be found in Supplementary Table S3. Of the 1881 precursors from miRBase v21, 321 are consistently detected in at least 50% of the samples for each biogroup. Among these, 282 were differentially expressed between PT and STN according to the FDR-adjusted wilcoxon test P-values ($P < 0.01$). Overrepresentation analysis of the precursors resulted in 541 significantly enriched and seven significantly depleted (FDR-adjusted; $P < 0.05$) categories. As shown in Figure 2A, a subset of precursors is ubiquitously present in significant categories, while others seem to be more specific. The top 10 categories sorted by increasing P-value are associated with cancer, including renal cell carcinoma. Also, the observed over expected ratio (123/48.6) indicates a strong enrichment ($P = 2.80 \times 10^{-38}$) of the de-regulated precursors with kidney and other types of cancer. A miRNA set enrichment analysis, using the list of detected precursors and sorted by effect size, revealed 253 enriched and 40 depleted categories. Here, the miRNA gene cluster 147, 189, 704 : 147, 284, 728 on the X chromosome is the most depleted category ($P = 8.64 \times 10^{-10}$), an observation that is in line with the depletion of precursor family hsa-mir-506. Interestingly, the list of highly enriched terms contains many transcription factors, the top 5 being *HEY1*, *WDR5*, *ELF1*, *BRD4* and *FLI1*.

Case study 2: mouse model for breast cancer progression

To showcase the novel support for model organisms in miEAA, we selected a dataset from GEO where circulating miRNAs from a breast-cancer mouse model were measured with microarrays (41). The dataset comprises 36 samples from mutation-carrier (NeuT+) and age-matched wildtype (NeuT–) mice that were collected at the premalignant, preinvasive and invasive stages of the disease. In this particular study, agilent microarrays probed with miRNAs from miRBase v19 were used on mice's plasma extracted RNA samples. Sample information can be found in Supplementary Table S4. Following a detection threshold procedure similar to our first case study, 212 miRNAs remained for differential expression analysis. Of these, mmu-miR-6243 had to be discarded as a result of mapping the identifiers from miRBase v19 to v22.1, which we performed with the miEAA miRBase version converter. Subsequently, we applied GSEA on the list of miRNAs sorted by decreasing effect size between the premalignant and the invasive stage, for NeuT+ and NeuT- samples separately. Strikingly, the former run returned 311 significant categories, while the latter returned none. Overall, many more categories seemed to be depleted ($N = 301$) than enriched ($N = 9$), suggesting a wide-spread up-regulation of molecular pathways as miRNAs get down-regulated in NeuT+. For example, we found Macrophage differentiation ($P = 2.54 \times 10^{-5}$), Vasculature development ($P = 1.60 \times 10^{-4}$), and VEGF signaling pathway ($P = 0.0016$) to be depleted, which might be a signal for the increased tumor burden of NeuT+ mice at the invasive breast cancer stage. Moreover, we evaluated GSEAs for the comparison of NeuT+ and NeuT- at all three stages. While the first two setups returned a rather unspecific set of categories with all P-values located close to the significance boundary, the last comparison yielded many interesting results. First, observations were in line with the group-wise comparison along the age dimension, because all categories are depleted, i.e. no enrichments at the top of the sorted list. Further, the results show that several dozen conserved miRNAs ($P = 4.53 \times 10^{-5}$) are down-regulated in the NeuT+ model at the invasive stage. More significant categories we



Step 1 of 5: Select miRNA or precursor as input for analysis

Input type: ● miRNA ○ miRNA precursor

Next

Step 2 of 5: Select type of enrichment analysis

Analysis type: ● Over-Representation Analysis (ORA) ○ miRNA enrichment analysis ((G)SEA)

Next

Step 3 of 5: Paste or upload testset

Paste example

Species: ● Homo sapiens ○ Mus musculus ○ Rattus norvegicus ○ Arabidopsis thaliana ○ Bos taurus ○ Caenorhabditis elegans ○ Drosophila melanogaster ○ Danio rerio ○ Gallus gallus ○ Sus scrofa

Testset:
hsa-miR-20b-5p
hsa-miR-144-5p
hsa-miR-17-5p
hsa-miR-20a-5p
hsa-miR-222-3p
hsa-miR-106a-5p
hsa-miR-93-5p
hsa-miR-126-3p
hsa-miR-363-3p
hsa-miR-302c-3p

Testset file: Browse... No file selected.

? Plain text file containing one miRNA or precursor per line
→ Example File miRNAs
→ Example File precursors

Next

Step 4 of 5: Paste or upload reference set (optional)

Reference set:

Reference set file: Browse... No file selected.

? Plain text file containing one miRNA or precursor per line
→ Example File miRNAs
→ Example File precursors

Next

Step 5 of 5: Select additional parameters for analysis

☑ Un/check all default categories

Categories:
☑ Annotations derived over miRTarBase (Gene Ontology)
☑ Pathways (KEGG)
☑ exRNA forms (miRandola)
☐ Chromosomal location (miRBase)
☑ Conservation (miRBase)
☑ Confidence (miRBase)
☑ Seed family (miRBase)
☐ Target genes (miRTarBase)
☑ Diseases (miRWalk)
☑ Gene Ontology (miRWalk)
☑ Organs (miRWalk)
☑ Pathways (miRWalk)
☑ Diseases (MNDR)
☑ Gender/Age
☐ Cell-type specific (Atlas)
☐ Published Diseases
☐ Immune cells
☑ Localization (RNALocate)
☑ Drugs (SM2miR)
☑ Expressed in tissue (Tissue Atlas)

Specialized categories (Expert mode): Show categories

P-value adjustment method:
○ No adjustment
● FDR (Benjamini-Hochberg) adjustment
○ Bonferroni adjustment
○ Benjamini-Yekutieli adjustment
○ Hochberg adjustment
○ Holm adjustment
○ Hommel adjustment

Benjamini-Hochberg FDR adjustment is also reported as 'Q-value' in results.

P-value adjustment scope:
● Adjust p-values for each category independently
○ Adjust p-values for all categories collectively

Significance level: 0.05

Minimum required hits per sub-category: 2

Submit

Results for input list

- View significant annotations for the uploaded miRNAs/precursors
- View sequence properties for the uploaded miRNAs/precursors

Show 5 rows CSV Excel

Category	Subcategory		Enrichment	P-value	P-adjusted	Q-value	Expected	Observed	miRNAs/precursors
				From To	From To	From To	From To	From To	
Annotations derived over miRTarBase (Gene Ontology)	U2 snRNA binding GO0030620	↑	over-represented	3.85e-9	5.46e-6	5.46e-6	1.12061	11	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-222-3p; hsa-miR-106a-5p; hsa-miR-93-5p; hsa-miR-374b-5p; hsa-miR-484; hsa-miR-106b-5p; hsa-miR-93-3p; hsa-miR-409-3p
Annotations derived over miRTarBase (Gene Ontology)	UDP-glucuronate decarboxylase activity GO0048040	↑	over-represented	5.34e-9	5.46e-6	5.46e-6	0.0901639	5	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-93-5p; hsa-miR-106b-5p
Annotations derived over miRTarBase (Gene Ontology)	cap2 mRNA methylation GO0097310	↑	over-represented	5.34e-9	5.46e-6	5.46e-6	0.0901639	5	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-93-5p; hsa-miR-106b-5p
Annotations derived over miRTarBase (Gene Ontology)	endothelial cell chemotaxis to vascular endothelial growth factor GO0090668	↑	over-represented	4.57e-9	5.46e-6	5.46e-6	0.167447	6	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-93-5p; hsa-miR-106b-5p; hsa-miR-409-3p
Annotations derived over miRTarBase (Gene Ontology)	eukaryotic 48S preinitiation complex GO0033290	↑	over-represented	3.59e-9	5.46e-6	5.46e-6	0.618267	9	hsa-miR-20b-5p; hsa-miR-17-5p; hsa-miR-20a-5p; hsa-miR-222-3p; hsa-miR-106a-5p; hsa-miR-93-5p; hsa-miR-302c-3p; hsa-miR-130b-5p; hsa-miR-106b-5p

Figure 1. miEAA workflow and exemplary results. (**A**) Each miRNA/precursor enrichment analysis consists of at most five steps. First, users should select whether they want to perform enrichment on precursors or miRNAs. Second, the enrichment algorithm, i.e. either ORA or GSEA must be selected. Next, the desired test set can be defined either through a textbox or a file upload. The fourth step only appears for ORAs where custom background reference sets can be inserted or uploaded. This is optional since miEAA provides pre-computed reference sets for all categories. Lastly, the set of categories and databases as well as statistical parameters should be selected. (**B**) Typical result view for an ORA. Users can sort, select, filter, and export the obtained enrichment results interactively. Moreover, several visualizations of the results are provided for each run, such as the precursor/miRNA to category heatmap and the category word cloud.

found such as exosome ($P = 2.31 \times 10^{-5}$) and circulating ($P = 0.0086$) miRNAs, breast cancer ($P = 0.0094$, Figure 2(b)), microRNAs in cancer ($P = 0.028$), and PI3K-Akt signaling pathway ($P = 0.028$) can be associated with the research setup of this exemplifying study.

Case study 3: Parkinson's Biomarkers from PPMI and comparison to TAM 2.0

At last we aimed to test a non-cancer disease (Parkinson's), to present a direct comparison between TAM 2 and miEAA 2.0. We compared the raw P-values of the tools to exclude an influence of the size of available categories. A direct comparison highlighted 72 hits by both tools (additional 70 reported only by TAM and 144 only by miEAA). Very similar but not exactly matching category names (e.g. *Alzheimer's* versus *Alzheimers* or *Carcinoma, Lung, Non-Small-Cell* versus *Carcinoma, Lung. Non-Small-Cell*) had to be matched manually. After matching those, several ambiguously defined categories remained, e.g. *Human Immunodeficiency Virus Infection* in miEAA and *Acquired Immunodeficiency Syndrome* in TAM and that had to be mapped. As a result, the overlap increased to 94 hits. Asking whether the overlap between the output of the two tools is larger for the categories with higher significance than expected, we performed a DynaVenn analysis of the result sets ordered by increasing P-value (42). Selecting the 32 most significant miEAA sets and the 30 most significant TAM sets we observed an overlap of 23 categories ($P = 10^{-8}$), indeed suggesting better comparable results for the most significant categories. Also, when comparing the miRNA hits for the obtained categories we observed very similar results. Alzheimer's Disease was covered by 10 miRNAs in miEAA and nine in TAM with P-values of 3.31×10^{-4} and 2.19×10^{-3}, respectively. We also observed the function category of TAM to be advantageous in this case, revealing direct hits such

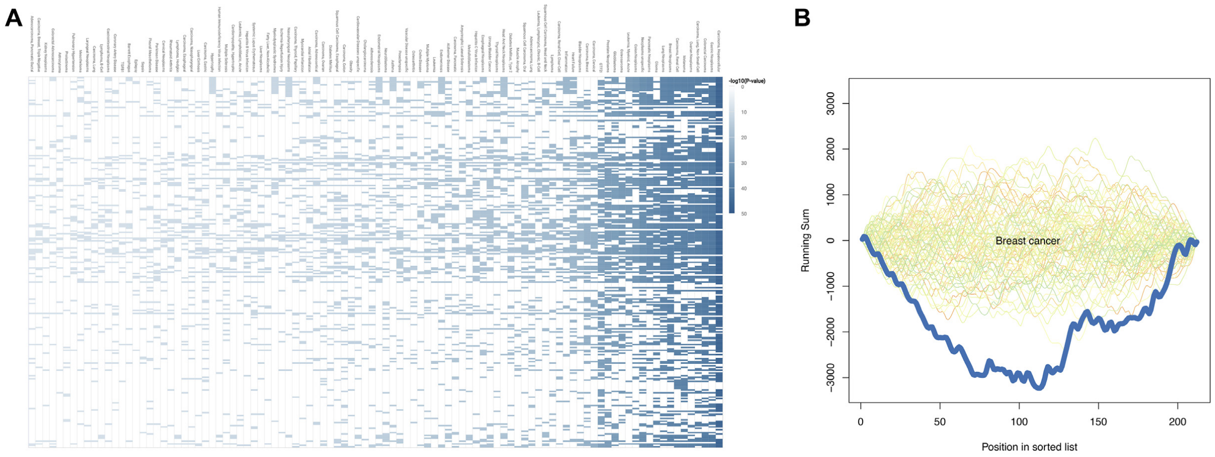


Figure 2. Web server visualisation of case study results. (**A**) Category (x-axis) to precursor (y-axis) heatmap with $-\log_{10}$-scaled enrichment *P*-values for the first case study. (**B**) GSEA plot with simulated background distributions (green to orange lines) and actual depletion for breast cancer (dark blue line) observed during evaluation of the second case study.

as *Aging*, which remained partially hidden in miEAA. On the other hand, miEAA seems to have slight advantages in the disease-associated categories, reporting 176 entries compared to 106 in TAM. This extended list contains among others *Parkinson's Disease* which was covered by three miRNAs in TAM and missing the alpha level while being covered by six miRNAs in miEAA and thus being significant ($P = 0.019$). The full list of results obtained from both tools in direct comparison is shown in Supplementary Table S5. Besides the case study benchmark, we performed a detailed feature comparison with respect to 22 criteria between our tool and TAM that is shown in Supplementary Table S6.

New data export and browsable API

All data, results, and interactive plots shown on the web server are exportable to common data formats. To support the trend towards the development of reproducible and automated data analysis pipelines (12), miEAA hosts a public, browsable API offering the same functionality as the web site, allowing one to access the miRNA converters and statistical algorithms remotely. This functionality is further augmented by a full-featured Python package with API library code and a command-line interface (CLI). For example, a regular workflow as performed on the website can be accomplished with three sequential calls to the web API or one call to the CLI. We provide code examples in the common data science programming languages Python and R to demonstrate this use-case. We also implemented the interface to solve two recurring problems in biological data analysis. First, reproducibility of statistical experiments can be improved, because usage of the versioned API in the context of a workflow manager such as Snakemake (26) or Nextflow fosters self-documenting research setups (43). Second, oftentimes the analysis of miRNA high-throughput data involves the comparison of multiple biogroups, timepoints or other annotation variables. By using our API and the package, multiple runs of miEAA can be performed at ease while minimising the time spent for set up and results aggregation.

DISCUSSION

Statistical tools for biological enrichment analysis are a key to understanding data from high-throughput omics assays. However, the performance primarily depends on the quality of the underlying annotations and the statistical soundness. We show that new developments in the miRNA research field yielded an unprecedented set of biological categories, covering most aspects of miRNA properties and function, with cross-species analysis becoming increasingly important. On the other side, as with every statistical framework applied on biological data, assumptions are not always met and findings should be assessed critically in the light of further validation experiments. The novel release of miEAA attempts to cover these aspects by enhancing the set of available categories both quantitatively and qualitatively as well as through offering more (stringent) approaches for *P*-value correction. Also, a major limitation of some datasets concerns the availability of mature miRNA identifiers, as only precursor names were available for some of the sources. However, especially in the context of diseases, mature miRNA resolution is preferable to match the biological selectivity for one major miRNA arm being expressed. Datasets incorporated in miEAA were compiled either automatically or manually. The competitor tool TAM uses a fewer number of high-quality annotations. In particular, an advantage of TAM arises from the manual curation of datasets (14). The case study on Parkinson's disease highlighted the results of miEAA 2.0 and TAM 2.0 to be similar whereas individual advantages in usability, functionality, or scope in the one or the other tool remain.

We have demonstrated the capability of miEAA to yield novel biological results in cancer research. For the kidney renal clear cell carcinoma case study, we found a depletion of the mir-506 precursor family, which has been observed before in other types of cancers (44,45). Many interactions to transcription factors were also found for the up-regulated miRNAs, suggesting an increased regulatory burden due to the exceeding transcriptional up-regulation observed in cancer. For example, HEY1, which is a transcriptional repressor has been characterised to be up-regulated in renal cell carcinomas (46). For the mouse breast cancer progression study, we illustrated the backwards compatibility of miEAA with respect to miRBase. The overall observed depletion of pathway regulating miRNAs in mice agrees with our first case study. Moreover, the significant categories like vasculature development that are associated with morphogenesis, resemble an increased tumor burden of NeuT+ mice, which was previously confirmed with a large human RNA-seq dataset on breast cancer (47). In both case studies, we observed many associations with other types of cancers or diseases. While this may speak for a molecular and biological similarity, a certain publication bias, e.g. for cancer, is a confounding factor that skews the statistics (14).

Establishing a standardized nomenclature is an on-going challenge in miRNA research. Results of the implemented manual converters are more accurate as compared to automated mappings since the naming schemes changed along the different releases. miEAA supports an exact mapping of old (e.g. miR*) to new nomenclature which would be ambiguous using automatic conversion (e.g. hsa-miR-499a-3p could be converted to hsa-miR-499a-3p or hsa-miR-499b-3p). Similar ambiguity issues would arise by performing a case insensitive miRNA to precursor mapping (‘miR’ to ‘mir’), in case multiple precursors with the same miRNA exist (for example hsa-let-7a-5p is annotated in three precursors). Finally, we sought to improve accessibility of miEAA and developed a web-API in combination with a Python package. These features enhance its usability in other applications for miRNA research, for example to annotate functional sub-graphs in regulatory network analysis (48). In conclusion, miEAA 2.0 is a flexible, comprehensive, and highly accessible tool for high-throughput miRNA annotation and enrichment analysis.

DATA AVAILABILITY

miEAA 2.0 is freely available at https://www.ccb.uni-saarland.de/mieaa2. No login is required. Example code for API-usage and the pre-compiled Python package are freely available from https://github.com/Xethic/miEAA-API.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the authors of the utilized GEO dataset for providing their microarray samples to the general public. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We would like to express gratitude towards all specimen donors and research groups involved in the sample acquisition.

FUNDING

Michael J. Fox foundation [14446]. Internal funds of Saarland University. Funding for open access charge: Michael J. Fox Foundation for Parkinson’s Research.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2018) Metazoan MicroRNAs. *Cell*, **173**, 20–51.
2. Kern,F., Backes,C., Hirsch,P., Fehlmann,T., Hart,M., Meese,E. and Keller,A. (2019) What’s the target: understanding two decades of in silico microRNA-target prediction. *Brief. Bioinform*, doi:10.1093/bib/bbz111.
3. Cantini,L., Bertoli,G., Cava,C., Dubois,T., Zinovyev,A., Caselle,M., Castiglioni,I., Barillot,E. and Martignetti,L. (2019) Identification of microRNA clusters cooperatively acting on epithelial to mesenchymal transition in triple negative breast cancer. *Nucleic Acids Res.*, **47**, 2205–2215.
4. Ludwig,N., Fehlmann,T., Kern,F., Gogol,M., Maetzler,W., Deutscher,S., Gurlit,S., Schulte,C., von Thaler,A.K., Deuschle,C. *et al.* (2019) Machine learning to detect Alzheimer’s disease from circulating Non-coding RNAs. *Genomics Proteomics Bioinformatics*, **17**, 430–440.
5. Thomou,T., Mori,M.A., Dreyfuss,J.M., Konishi,M., Sakaguchi,M., Wolfrum,C., Rao,T.N., Winnay,J.N., Garcia-Martin,R., Grinspoon,S.K. *et al.* (2017) Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature*, **542**, 450–455.
6. Backes,C., Meese,E. and Keller,A. (2016) Specific miRNA disease biomarkers in blood, serum and Plasma: Challenges and prospects. *Mol. Diagn. Ther.*, **20**, 509–518.
7. Fehlmann,T., Backes,C., Alles,J., Fischer,U., Hart,M., Kern,F., Langseth,H., Rounge,T., Umu,S.U., Kahraman,M. *et al.* (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, **34**, 1621–1628.
8. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grässer,F.A., Lenhof,H.P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, **47**, 3353–3364.
9. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
10. Mora,A. (2019) Gene set analysis methods for the functional interpretation of non-mRNA data—genomic range and ncRNA data. *Brief. Bioinform*, doi:10.1093/bib/bbz090.
11. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) MiEAA: MicroRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.
12. Perkel,J.M. (2019) Workflow systems turn raw data into scientific knowledge. *Nature*, **573**, 149–150.
13. Lu,M., Shi,B., Wang,J., Cao,Q. and Cui,Q. (2010) TAM: A method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics*, **11**, 419.
14. Li,J., Han,X., Wan,Y., Zhang,S., Zhao,Y., Fan,R., Cui,Q. and Zhou,Y. (2018) TAM 2.0: Tool for MicroRNA set analysis. *Nucleic Acids Res.*, **46**, W180–W185.
15. Çorapçıoğlu,M. and Oğul,H. (2015) miSEA: microRNA set enrichment analysis. *Biosystems*, **134**, 37–42.
16. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
17. Backes,C., Fehlmann,T., Kern,F., Kehl,T., Lenhof,H.P., Meese,E. and Keller,A. (2018) MiRCarta: A central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.
18. Huang,H.Y., Lin,Y.C.D., Li,J., Huang,K.Y., Shrestha,S., Hong,H.C., Tang,Y., Chen,Y.G., Jin,C.N., Yu,Y. *et al.* (2019) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.*, **48**, D148–D154.
19. Kehl,T., Kern,F., Backes,C., Fehlmann,T., Stöckel,D., Meese,E., Lenhof,H.P. and Keller,A. (2020) miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res.*, **48**, D142–D147.
20. Ludwig,N., Leidinger,P., Becker,K., Backes,C., Fehlmann,T., Pallasch,C., Rheinheimer,S., Meder,B., Stähler,C., Meese,E. *et al.* (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res.*, **44**, 3865–3877.

21. Huang,Z., Shi,J., Gao,Y., Cui,C., Zhang,S., Li,J., Zhou,Y. and Cui,Q. (2018) HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.*, **47**, D1013–D1017.
22. Cui,T., Zhang,L., Huang,Y., Yi,Y., Tan,P., Zhao,Y., Hu,Y., Xu,L., Li,E. and Wang,D. (2017) MNDR v2.0: an updated resource of ncRNA–disease associations in mammals. *Nucleic Acids Res.*, **46**, D371–D374.
23. Tong,Z., Cui,Q., Wang,J. and Zhou,Y. (2018) TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Res.*, **47**, D253–D258.
24. Zhang,T., Tan,P., Wang,L., Jin,N., Li,Y., Zhang,L., Yang,H., Hu,Z., Zhang,L., Hu,C. *et al.* (2016) RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.*, **45**, D135–D138.
25. Russo,F., Di Bella,S., Vannini,F., Berti,G., Scoyni,F., Cook,H.V., Santos,A., Nigita,G., Bonnici,V., Laganà,A. *et al.* (2017) miRandola 2017: a curated knowledge base of non-invasive biomarkers. *Nucleic Acids Res.*, **46**, D354–D359.
26. Köster,J. and Rahmann,S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
27. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J. *et al.* (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
28. The Gene Ontology Consortium (2018) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
29. Kanehisa,M., Sato,Y., Furumichi,M., Morishima,K. and Tanabe,M. (2018) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
30. Teng,X., Chen,X., Xue,H., Tang,Y., Zhang,P., Kang,Q., Hao,Y., Chen,R., Zhao,Y. and He,S. (2019) NPInter v4.0: an integrated database of ncRNA interactions. *Nucleic Acids Res.*, **48**, D160–D165.
31. Liu,X., Wang,S., Meng,F., Wang,J., Zhang,Y., Dai,E., Yu,X., Li,X. and Jiang,W. (2012) SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics*, **29**, 409–411.
32. de Rie,D., Abugessaisa,I., Alam,T., Arner,E., Arner,P., Ashoor,H., Åström,G., Babina,M., Bertin,N., Burroughs,A.M. *et al.* (2017) An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.*, **35**, 872–878.
33. Minami,K., Uehara,T., Morikawa,Y., Omura,K., Kanki,M., Horinouchi,A., Ono,A., Yamada,H., Ohno,Y. and Urushidani,T. (2014) miRNA expression atlas in male rat. *Scientific Data*, **1**, 140005.
34. Dweep,H. and Gretz,N. (2015) MiRWalk2.0: A comprehensive atlas of microRNA-target interactions. *Nat. Methods*, **12**, 697.
35. The RNAcentral Consortium (2018) RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.*, **47**, D221–D229.
36. Xu,T., Su,N., Liu,L., Zhang,J., Wang,H., Zhang,W., Gui,J., Yu,K., Li,J. and Le,T.D. (2018) miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. *BMC Bioinformatics*, **19**, 514.
37. Keller,A., Backes,C. and Lenhof,H.P. (2007) Computation of significance scores of unweighted gene set enrichment analyses. *BMC Bioinformatics*, **8**, 290.
38. Marek,K., Chowdhury,S., Siderowf,A., Lasch,S., Coffey,C.S., Caspell-Garcia,C., Simuni,T., Jennings,D., Tanner,C.M., Trojanowski,J.Q. *et al.* (2018) The Parkinson's progression markers initiative (PPMI)– establishing a PD biomarker cohort. *Ann. Clin. Transl. Neur.*, **5**, 1460–1477.
39. Stöckel,D., Kehl,T., Trampert,P., Schneider,L., Backes,C., Ludwig,N., Gerasch,A., Kaufmann,M., Gessler,M., Graf,N. *et al.* (2016) Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, **32**, 1502–1508.
40. Korthauer,K., Kimes,P.K., Duvallet,C., Reyes,A., Subramanian,A., Teng,M., Shukla,C., Alm,E.J. and Hicks,S.C. (2019) A practical guide to methods controlling false discoveries in computational biology. *Genome. Biol.*, **20**, 118.
41. Chiodoni,C., Cancila,V., Renzi,T.A., Perrone,M., Tomirotti,A.M., Sangaletti,S., Botti,L., Dugo,M., Milani,M., Bongiovanni,L. *et al.* (2020) Transcriptional profiles and stromal changes reveal bone marrow adaptation to early breast cancer in association with deregulated circulating microRNAs. *Cancer Res.*, **80**, 484–498.
42. Amand,J., Fehlmann,T., Backes,C. and Keller,A. (2019) DynaVenn: web-based computation of the most significant overlap between ordered sets. *BMC Bioinformatics*, **20**, 743.
43. DI Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
44. Li,J., Wu,H., Li,W., Yin,L., Guo,S., Xu,X., Ouyang,Y., Zhao,Z., Liu,S., Tian,Y. *et al.* (2016) Downregulated miR-506 expression facilitates pancreatic cancer progression and chemoresistance via SPHK1/Akt/NF-κB signaling. *Oncogene*, **35**, 5501–5514.
45. Zhang,L., Zhou,H. and Wei,G. (2019) miR-506 regulates cell proliferation and apoptosis by affecting RhoA/ROCK signaling pathway in hepatocellular carcinoma cells. *Int. J. Clin. Exp. Pathol.*, **12**, 1163–1173.
46. Karim,S., Al-Maghrabi,J.A., Farsi,H.M.A., Al-Sayyad,A.J., Schulten,H.J., Buhmeida,A., Mirza,Z., Al-boogmi,A.A., Ashgan,F.T., Shabaad,M.M. *et al.* (2016) Cyclin D1 as a therapeutic target of renal cell carcinoma- a combined transcriptomics, tissue microarray and molecular docking study from the Kingdom of Saudi Arabia. *BMC Cancer*, **16**, 741.
47. Tapia-Carrillo,D., Tovar,H., Velazquez-Caldelas,T.E. and Hernandez-Lemus,E. (2019) Master regulators of signaling Pathways: An application to the analysis of gene regulation in breast cancer. *Front. Genet.*, **10**, 1180.
48. Fan,Y., Siklenka,K., Arora,S.K., Ribeiro,P., Kimmins,S. and Xia,J. (2016) miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res.*, **44**, W135–W141.

3.23 *Machine learning to detect Alzheimer's disease from circulating non-coding RNAs*

Genomics Proteomics Bioinformatics 17 (2019) 430–440

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Machine Learning to Detect Alzheimer's Disease from Circulating Non-coding RNAs

Nicole Ludwig [1,#,a], Tobias Fehlmann [2,#,b], Fabian Kern [2,#,c], Manfred Gogol [3,d], Walter Maetzler [4,5,6,e], Stephanie Deutscher [1,f], Simone Gurlit [7,g], Claudia Schulte [5,6,h], Anna-Katharina von Thaler [5,6,i], Christian Deuschle [5,6,j], Florian Metzger [8,k], Daniela Berg [4,5,6,l], Ulrike Suenkel [5,6,m], Verena Keller [9,n], Christina Backes [2,o], Hans-Peter Lenhof [10,p], Eckart Meese [1,q], Andreas Keller [2,10,*,r]

[1] *Department of Human Genetics, Saarland University, 66421 Homburg/Saar, Germany*
[2] *Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany*
[3] *Institut für Gerontologie, Universität Heidelberg, 69047 Heidelberg, Germany*
[4] *Department of Neurology, Christian-Albrechts-University of Kiel, 24118 Kiel, Germany*
[5] *Center for Neurology and Hertie Institute for Clinical Brain Research, Department of Neurodegeneration, University of Tuebingen, 72074 Tuebingen, Germany*
[6] *German Center for Neurodegenerative Diseases (DZNE), 72076 Tuebingen, Germany*
[7] *Department of Anesthesiology and Intensive Care, St. Franziskus Hospital Muenster, 48145 Muenster, Germany*
[8] *Department of Psychiatry and Psychotherapy, University Hospital Tuebingen, 72016 Tuebingen, Germany*
[9] *Department of Medicine II, Saarland University Medical Center, 66421 Homburg/Saar, Germany*
[10] *Center for Bioinformatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany*

[*] Corresponding author.
E-mail: andreas.keller@ccb.uni-saarland.de (Keller A).

[#] Equal contribution.

[a] ORCID: 0000-0003-4703-7567.
[b] ORCID: 0000-0003-1967-2918.
[c] ORCID: 0000-0002-8223-3750.
[d] ORCID: 0000-0001-7169-2970.
[e] ORCID: 0000-0002-5945-4694.
[f] ORCID: 0000-0003-4080-708X.
[g] ORCID: 0000-0002-3944-7841.
[h] ORCID: 0000-0003-4006-1265.
[i] ORCID: 0000-0001-8161-5813.
[j] ORCID: 0000-0001-5571-7293.
[k] ORCID: 0000-0002-8236-1170.
[l] ORCID: 0000-0003-1187-219X.
[m] ORCID: 0000-0002-5348-3996.
[n] ORCID: 0000-0003-3240-1397.
[o] ORCID: 0000-0001-9330-9290.
[p] ORCID: 0000-0002-5820-9961.
[q] ORCID: 0000-0001-7569-819X.
[r] ORCID: 0000-0002-5361-0895.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2019.09.004

Received 14 May 2019; revised 26 August 2019; accepted 7 October 2019
Available online 4 December 2019

Handled by Hongxing Lei

KEYWORDS

miRNAs;
Neurodegeneration;
Alzheimer's disease;
Biomarker;
Non-coding RNAs;
Gene regulation

Abstract Blood-borne small non-coding (sncRNAs) are among the prominent candidates for blood-based diagnostic tests. Often, high-throughput approaches are applied to discover **biomarker** signatures. These have to be validated in larger cohorts and evaluated by adequate statistical learning approaches. Previously, we published high-throughput sequencing based microRNA (miRNA) signatures in **Alzheimer's disease** (AD) patients in the United States (US) and Germany. Here, we determined abundance levels of 21 known circulating **miRNAs** in 465 individuals encompassing AD patients and controls by RT-qPCR. We computed models to assess the relation between miRNA expression and phenotypes, gender, age, or disease severity (Mini-Mental State Examination; MMSE). Of the 21 miRNAs, expression levels of 20 miRNAs were consistently de-regulated in the US and German cohorts. 18 miRNAs were significantly correlated with **neurodegeneration** (Benjamini-Hochberg adjusted $P < 0.05$) with highest significance for miR-532-5p (Benjamini-Hochberg adjusted $P = 4.8 \times 10^{-30}$). Machine learning models reached an area under the curve (AUC) value of 87.6% in differentiating AD patients from controls. Further, ten miRNAs were significantly correlated with MMSE, in particular miR-26a/26b-5p (adjusted $P = 0.0002$). Interestingly, the miRNAs with lower abundance in AD were enriched in monocytes and T-helper cells, while those up-regulated in AD were enriched in serum, exosomes, cytotoxic t-cells, and B-cells. Our study represents the next important step in translational research for a miRNA-based AD test.

Introduction

Alzheimer's disease (AD) represents one of the most demanding challenges in healthcare [1,2]. In light of demographic changes and failures in drug development [3], early detection of the disease offers itself as one of the most promising approaches to improve patients' outcome in the mid- to long term. Especially minimally invasive molecular markers seem to have a significant potential to facilitate a diagnosis of AD, even in early stages.

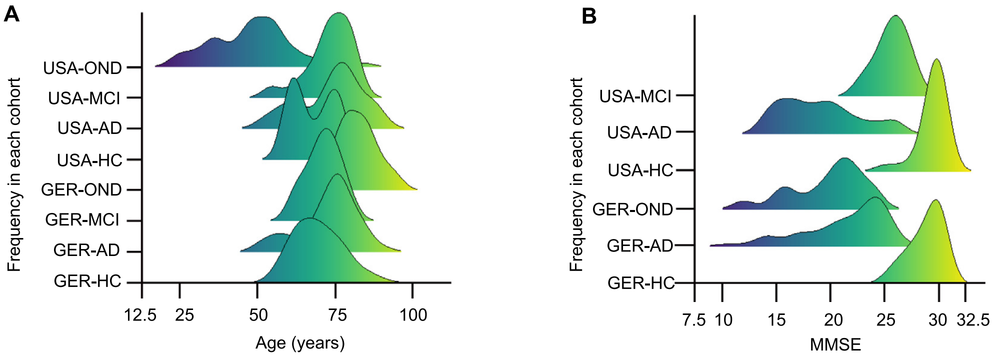
The importance of minimally invasive molecular markers for AD is reflected by over 3000 original articles and reviews related to AD diagnosis from blood, serum, or plasma samples published and indexed in PubMed. Among the promising approaches are plasma proteomic markers measured by mass spectrometry [4], metabolic patterns [5], gene expression profiles [6], DNA methylation [7], and small non-coding RNAs (sncRNAs) [8]. However, cohort sizes of such studies are often limited and larger validation cohorts frequently did not always match the original results [9]. One of the major challenges is the complexity of signatures that is often required to reach high specificity and sensitivity.

For AD, many miRNA-related studies from tissue [10], blood [11], serum [12], exosomes, [13] or cerebrospinal fluid (CSF) [12] have been performed. In one of the most comprehensive reviews [14], Hu and co-workers investigated 236 papers and reviewed the de-regulated miRNA abundance in different parts of AD patients. In another comprehensive recent review, Nagaraj and co-workers show that out of 137 miRNAs found to exhibit altered expression in AD blood, 36 have been replicated in at least one independent study. Moreover, out of 166 miRNAs being differentially abundant in AD CSF, 13 have been repeatedly found [15].

In previous studies, we performed deep sequencing to measure blood-borne AD miRNA signatures in a cohort of 54 AD patients and 22 controls from the United States (USA) that have been partially validated on a larger cohort of 202 samples by RT-qPCR [8]. In a second study using the same technique, we aimed to validate the results in a patient cohort collected in Germany (GER) that included 49 AD cases, 55 controls and 110 disease controls [16]. The results of both studies were largely consistent with a correlation between both studies of 0.93 (95% confidence interval 0.89–0.96; $P < 10^{-16}$).

Although deep-sequencing applications are increasingly introduced into clinical care, they are mostly performed for the analysis of DNA or RNAs coding for genes. Small non-coding RNA profiling, however, is mostly achieved by microarray and RT-qPCR based approaches. In the present study, we provide further evidence that blood-borne miRNA signatures can be measured by standard RT-qPCR, becoming valuable tools for the minimally-invasive detection of AD. From our above-mentioned studies and the literature, we selected a set of 21 miRNAs and determined the abundance of these miRNAs in the blood of 465 individuals. The 465 individuals consist of 169 individuals from our initial study (36%) [8], 107 individuals from the second study (23%) [16] as well as 189 newly collected individuals (41%). An overview and summary on the German and US samples is provided in **Figure 1**A–C, the full details for each individual samples, including age gender, diagnosis, Mini-Mental State Examination (MMSE), and the miRNA measurements, are provided in Table S1.

With the present study we pursue the five main goals to demonstrate that (1) miRNAs from NGS studies can be well reproduced by RT-qPCR experiments; (2) given a reasonable heterogeneity in samples still reproducible measurements in



C

	AD		MCI		HC		OND	
	USA	GER	USA	GER	USA	GER	USA	GER
No of subjects	79	66	17	21	23	191	50	18
Gender (M/F)	40/39	36/30	8/9	10/11	11/12	92/99	16/34	7/11
Age (mean ± SD)	73.2 ± 10.3	72.5 ± 9.3	72.8 ±7.2	70.6 ± 5.3	67.4 ±6.7	69.1 ±8	48.2 ± 13.5	81.7 ± 6.4
MMSE (mean ± SD)	19 ± 3.5	21 ±4.7	25.6 ±1.4	NA	29.4 ± 1.2	28.8 ±1.5	NA	18.8 ± 5.3

Figure 1 Distribution of age, gender, diseases, and MMSE
A. Histogram for the age distribution in the different cohorts. The diagram shows for each cohort/disease the age distribution. Only the OND group from the US shows a deviation towards younger patients, while all other groups have similar age ranges. **B.** Histogram for the MMSE values. HCs and MCI patients show significantly larger MMSE values as compared to AD and OND patients. **C.** Metrics. For each of the cohorts and diseases, the number of patients in the US and Germany, the mean and SD for age and MMSE as well as the gender distribution are provided. GER, Germany; MMSE, Mini-Mental State Examination; AD, Alzheimer's disease; OND, other neurological diseases; HC, healthy control; MCI, mild cognitive impairment.

larger cohorts are possible; (3) miRNAs are also correlated to clinical features such as the MMSE value; (4) statistical learning approaches with as few as possible features lead to accurate diagnostic results; (5) the miRNAs likely have functionality in AD via targeting genes.

Results

Two endogenous control RNAs show concordant results

Because the selection of the most appropriate endogenous control RNAs for RT-qPCR experiments can be challenging, we previously evaluated systematically whether different endogenous controls lead to differences in miRNA measurements [17]. Especially, most miRNAs seem to be affected by development stages, tissues [18], or diseases [19], limiting their ability as controls and calling for endogenous controls other than miRNAs. Our results suggested that differences can be observed that are however moderate. In the present study we nonetheless evaluated and compared the performance of two commonly used endogenous controls RNU48 and RNU6. Both endogenous controls have been measured in duplicates. In comparing the results, we verified the generally high concordance between the two endogenous controls with a Pearson correlation of 0.854 (95% CI: 0.828–0.877; $P < 10^{-16}$). We thus report the result in the current study based on our standard endogenous control RNU48.

In the same direction we also investigated the general stability of RT-qPCR based miRNA measurement. One control sample has been measured 12 times over the study for all miRNAs. The median Pearson correlation coefficient (PCC) exceeded 0.99 as the heatmap and the box plot in Figure S1 show.

miRNAs are highly significantly correlated with neurodegeneration

In total, 465 participants have been analyzed by RT-qPCR. The abundance levels of 18 of the 21 miRNAs were significantly different between the four groups considered, *i.e.*, AD, mild cognitive impairment (MCI), other neurological diseases (OND), and healthy controls (HC). With an Benjamini-Hochberg (BH) adjusted P value of 4.8×10^{-30}, the most significant miRNA was miR-532-5p, which showed markedly decreased levels in AD patients, and slightly decreased levels in patients with OND and MCI (**Figure 2**A). The abundance levels of miR-17-3p, the miRNA with the second lowest P value ($P = 8.8 \times 10^{-28}$), showed a similar pattern as miR-532-5p (PCC > 0.9). The overall correlation matrix between the 21 miRNAs showed three large clusters of miRNAs with similar expression in the following referred to as Clusters A, B, and C (Figure 2B). The third and fourth most significant miRNAs in ANOVA, *i.e.*, miR-103a-3p and miR-107 ($P = 2.4 \times 10^{-18}$ and $P = 3.6 \times 10^{-15}$, respectively), came from Cluster C, like miR-532-5p, and miR-17-3p. MiR-1468-5p (Cluster A, $P = 6.2 \times 10^{-12}$; Figure 2C) shows an opposite expression pattern, i.e. a higher abundance in AD patients as compared to HC. The boxplots in Figure 2A/2C also underline that the deregulation of these miRNAs is strongest in AD compared to the HC. There is, however, a deregulation in MCI or OND, but to a lesser extent, such that the altered abundance is at least partially specific for AD. This result is consistent with our previous work based on high-throughput sequencing.

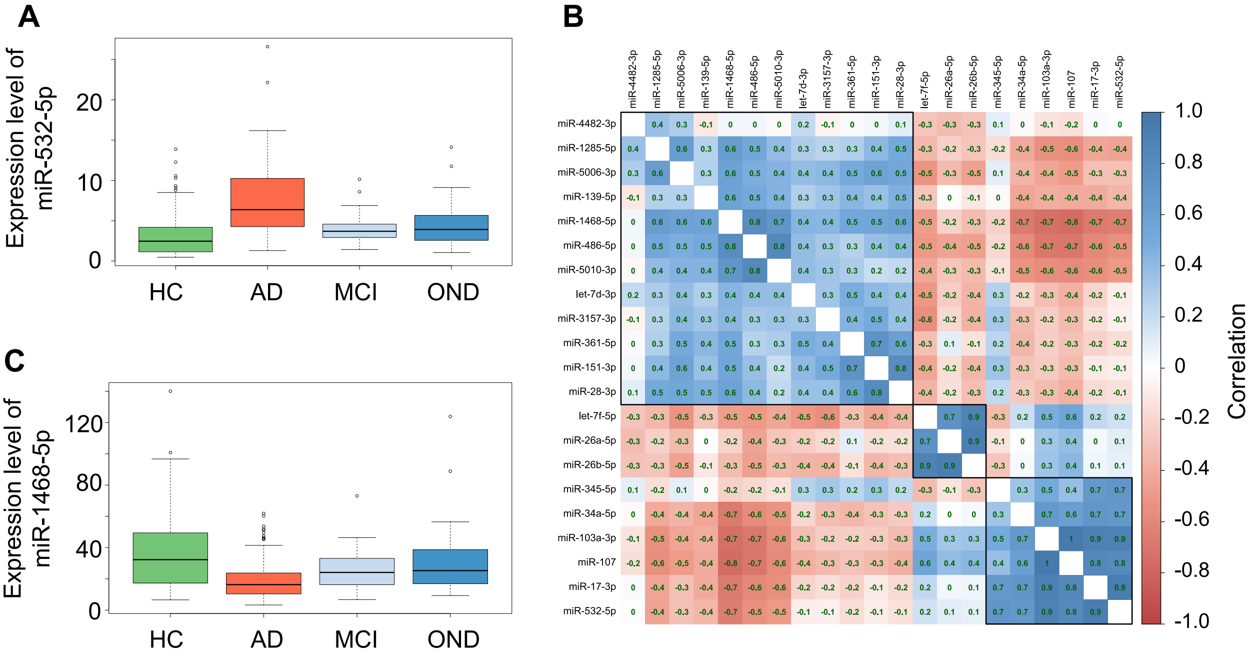


Figure 2 miRNAs are specifically dysregulated in the four cohorts and are partially co-expressed

A. Expression of miR-532-3p. The boxes display the 2nd and 3rd quartile of expression values for miR-532-3p in HC, patients with AD, MCI, or OND. The range of expression values in the four groups is indicated by the error bars with outliers represented by unfilled dots. Median expression of miR-532-3p is indicated as thick black line. **B.** Correlation of miRNA expression. This correlation matrix graphically represents the pair-wise correlation coefficient for all miRNAs tested. According to the color scale on the right side of the matrix, positive and negative correlations are indicated in shades of blue and red, respectively. PCC is given for each pair-wise correlation. Three clusters of miRNAs with highly similar expression patterns are indicated as Clusters A, B, and C on the left side. **C.** Expression of miR-1468-5p. The boxes display the 2nd and 3rd quartile of expression values for miR-1468-5p in HC, patients with AD, MCI, or OND. The range of expression values in the four groups is indicated by the error bars with outliers represented by unfilled dots. Median expression of miR-1468-5p is indicated as thick black line. PCC, Pearson correlation coefficient.

For a more detailed understanding of the miRNAs and their correlation to AD and other factors, we next assessed whether the abundance levels were correlated to age or gender, or, in case of AD and MCI with the MMSE results (Table 1). As Table 1 highlights, none of the miRNAs was associated with gender and five miRNAs were weakly associated with age of patients. Following adjustment for multiple testing, 14 miRNAs showed a significant differential expression in AD patients compared to controls (*i.e.*, HC, MCI, and OND combined). The above mentioned miR-532-5p and miR-17-3p were again the most significant markers for AD. Furthermore, ten miRNAs were significantly correlated with the MMSE value. Interestingly, all three miRNAs of Cluster B (Figure 1B), *i.e.*, miR-26a, 26b-5p, and let-7f-5p, showed the highest significance for the correlation to MMSE ($P < 0.005$). Since neither all miRNAs nor the MMSE values were normally distributed we repeated the analyses with non-parametric and ranked based Spearman correlation coefficient (SCC), overall leading to comparable results (see Table S2).

Besides the comparison of healthy controls to AD we also asked whether MCI patients can be separated from AD patients using miRNAs. Indeed, eleven miRNAs had significant differential expression in MCI versus AD following adjustment for multiple testing: miR-17-3p ($P = 10^{-12}$; down in AD), miR-532-5p ($P = 8 \times 10^{-10}$; down in AD), miR-103a-3p ($P = 10^{-8}$; down in AD), miR-107 ($P = 4 \times 10^{-7}$; down in AD), let-7d-3p ($P = 9 \times 10^{-7}$; up in AD), let-7f-5p ($P = 3 \times 10^{-5}$; down in AD), miR-345-5p ($P = 0.0002$; down in AD), miR-26a-5p ($P = 0.002$; down in AD), miR-26b-5p ($P = 0.009$; down in AD), miR-1468-5p ($P = 0.02$; up in AD), and miR-139-5p ($P = 0.03$; up in AD).

miRNA profiles from the US and German cohort show consistent results

It is essential to understand whether biomarkers can be concordantly determined in different cohorts. Although a direct comparison of ethnic groups was not in the scope of our analysis we nonetheless asked whether miRNA profiles for one disease measured on two different continents are concordant to each other. We thus compared the profiles measured from GER and USA cohorts. As the GER cohort was about twice as large as the USA cohort and P values depend on the number of individuals in each cohort, a comparison based only on P values is potentially biased. Therefore, we computed the fold changes (on a logarithmic scale) between AD and controls (Figure 3A). In this plot miRNAs in the upper right quadrant are down-regulated and miRNAs in the lower left quadrant

Table 1 Raw and adjusted *P* values of miRNAs for age, gender, AD, and MMSE

miRNA	Gender		Age		AD		MMSE	
	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
miR-**532**-5p	0.3466	0.4917	0.3089	0.4634	5.99E–22	**1.26E–20**	0.5048	0.5890
miR-**17**-3p	0.4885	0.5811	0.0639	0.2238	6.24E–18	**6.55E–17**	0.5004	0.5890
miR-**1468**-5p	0.0568	0.1491	0.5645	0.6973	5.98E–16	**4.19E–15**	0.0016	0.0057
miR-**5010**-3p	0.0176	0.0971	0.4787	0.6283	4.81E–12	**2.52E–11**	0.0131	0.0345
miR-**103**a-3p	0.3596	0.4917	0.2791	0.4508	1.56E–11	**6.56E–11**	0.5805	0.6416
miR-**1285**-5p	0.5195	0.5811	0.8097	0.8502	4.85E–11	**1.70E–10**	0.2269	0.3725
miR-**345**-5p	0.2217	0.4041	0.0008	0.0081	8.85E–11	**2.65E–10**	0.0174	0.0364
miR-**107**	0.2174	0.4041	0.3568	0.4995	6.94E–10	**1.82E–09**	0.7545	0.7545
miR-**486**-5p	0.5535	0.5811	0.9667	0.9667	2.79E–06	**6.51E–06**	0.2306	0.3725
miR-**139**-5p	0.0031	0.0656	0.7862	0.8502	8.12E–05	**0.0002**	0.3384	0.4441
miR-**361**-5p	0.3747	0.4917	0.0032	0.0180	0.0004	**0.0008**	0.0896	0.1710
miR-**5006**-3p	0.2271	0.4041	0.1433	0.3352	0.0006	**0.0011**	0.0043	0.0130
miR-**28**-3p	0.2309	0.4041	0.6780	0.7909	0.0057	0.0093	0.6572	0.6900
miR-**34**a-5p	0.0185	0.0971	0.2526	0.4508	0.0067	0.0100	0.2486	0.3730
miR-**4482**-3p	0.0378	0.1325	0.0004	0.0078	0.0365	0.0511	0.0015	0.0057
let-7f-5p	0.0454	0.1362	0.1596	0.3352	0.0702	0.0921	0.0002	**0.0016**
miR-**3157**-3p	0.3504	0.4917	0.2626	0.4508	0.0833	0.1029	0.0013	0.0057
miR-**151**-3p	0.0131	0.0971	0.1057	0.3170	0.0902	0.1052	0.3031	0.4244
miR-**26**b-5p	0.7003	0.7003	0.0034	0.0180	0.1101	0.1217	1.82E–05	**0.0002**
miR-**26**a-5p	0.5506	0.5811	0.0063	0.0263	0.2939	0.3085	1.19E–05	**0.0002**
let-7d-3p	0.0323	0.1325	0.1489	0.3352	0.4834	0.4834	0.0172	0.0364

Note: *P* values for gender and AD were calculated based on *t* test; *P* values for age and MMSE were calculated based on Pearson's product moment correlation coefficient. *P* values were adjusted by the Benjamini-Hochberg procedure. Adjusted *P* values < 0.05 are indicted in orange and those < 0.005 are put in bold with blue background. AD, Alzheimer's disease; MMSE, Mini-Mental State Examination.

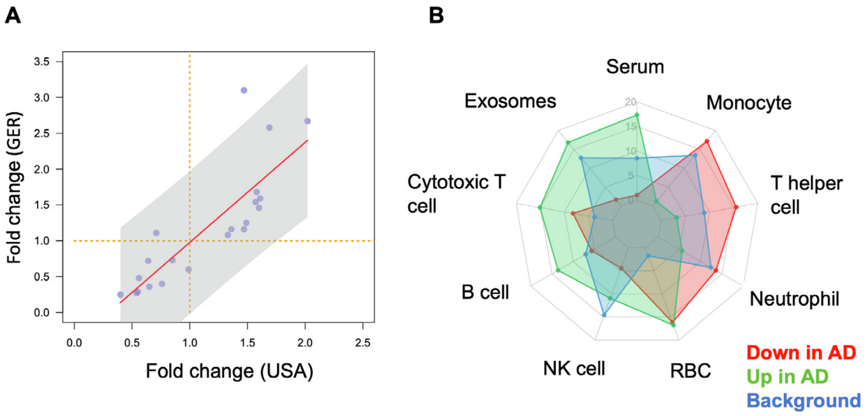


Figure 3 Differentially-expressed miRNAs are concordantly expressed in the German and the US cohorts and belong to specific blood compounds

A. Fold change in the USA cohort compared to the GER cohort. The X- and Y-axes represent the fold change between AD and HC on a $\log_2$ scale for the USA and GER patient cohorts, respectively. Each miRNA is represented by one dot. The dashed orange line is the segregation between up- and down-regulation. miRNAs in the upper right or lower left quadrant are concordantly up- or downregulated in AD compared to HC in both cohorts, respectively. The solid red line is a linear regression fit and the shaded area is the 95% confidence interval of that fit. **B.** Radar chart showing the blood compound distribution. The plot shows the relative abundance of up-regulated, down-regulated, and all miRNAs in different blood compounds. Since the relative abundance is provided, it is more appropriate to compare the different groups within one specific compound rather than comparing different compounds to each other.

are up-regulated in AD compared to controls concordantly in both cohorts. Of 21 miRNAs, only miR-4482-3p was down-regulated in the GER cohort, but up-regulated in the USA cohort. The differences in abundance levels of this miRNA in AD compared to controls were, however, not significant, neither in the GER nor in the USA cohort, nor in the combined analysis. Thus, miR-4482-3p likely represents a single false positive marker from the initial deep-sequencing based miRNA discovery study. In contrast, the results for the remaining 20 miRNAs were concordant between the USA and the GER cohort. Furthermore, eleven of these miRNAs were nominally significant in both cohorts, when analyzing the USA cohort and the GER cohort separately, and remained significant in the combined analysis. These significant miRNAs include miR-103a-3p, miR-107, miR-1285-5p, miR-139-5p, miR-1468-5p, miR-17-3p, miR-28-3p, miR-361-5p, miR-5006-3p, miR-5010-3p, and miR-532-5p.

Up- and down-regulated miRNAs are expressed in different blood compounds

We asked whether the miRNAs that are up- and down-regulated are expressed to the same amount in different blood cell types, serum or exosomes. To this end we made use of a public miRNA blood cell type atlas [20]. For the up- and down-regulated miRNAs we then compared the average expression in the different compounds and compared them to the background distribution of all human miRNAs (Figure 3B). Interestingly, we observed a highly specific pattern. miRNAs up-regulated in AD were expressed mostly in serum, exosomes, cytotoxic t-cells, and b-cells while those that were down-regulated in AD were expressed in monocytes and t-helper cells. These results suggest a complex regulatory pattern of miRNAs in the different blood cell compounds which would have been likely not observed if only a specific blood cell type or serum would have been investigated.

Machine learning facilitates accurate diagnosis of AD

To obtain more accurate diagnostic results, molecular markers can be considered as "weak learners" that can be combined by machine learning approaches. For our present data set, we explored common statistical and deep learning approaches including support vector machines, decision trees, neural networks and gradient boosted trees and others using five repeated runs of a ten-fold cross validation. While the performance of all approaches was similar (data not shown), the best results were obtained by gradient boosted trees. Compared to other classifiers, gradient boosted trees have the additional advantage that missing values do not have to be imputed. In the classification, two scenarios were modeled: First, the diagnosis of AD patients with unaffected controls (HC) as background group, and second, the diagnosis of AD patients with all controls, *i.e.*, HC, OND, and MCI combined, as background group. In the first and apparently less complex scenario the gradient boosted tree model reached an area under the curve (AUC) of 87.6% (**Figure 4**A). For the second and more complex case, an AUC of 83.5% was reached (Figure 4B). A further advantage of the gradient boosted tree models is that sensitivity and specificity can be well balanced and traded-off. Depending on whether a diagnosis trimmed for sensitivity or for specificity is required *e.g.*, in screening tests, as confirmatory tests or tests for enrollment for clinical studies, a sensitive or a specific model can be chosen.

Feature importance values for each miRNA based on the relative gain obtained via their splits were extracted from both models using the method provided by LightGBM (Table S3) According to this metric, miR-17-3p had the highest importance value in both models, followed by miR-5010-3p. For the model comparing AD to all controls, the next most important miRNAs were let-7d-3p, miR-26b-5p, and miR-28-3p. For the model comparing to unaffected controls, miR-361-5p, let-7d-3p, and miR-532-5p were the next most important features. Interestingly, let-7d-3p and miR-26b-5p were not significantly associated with AD on their own, suggesting that their discriminative power might come from the combination with other miRNAs or their association with different stages of the disease. For example, miR-26b-5p was recently reported to be likely deregulated early in AD, even before the appearance of clinical symptoms [21].

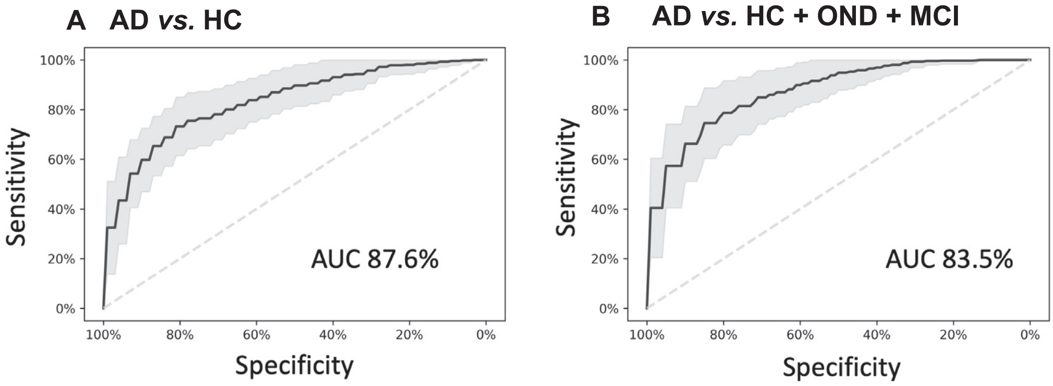
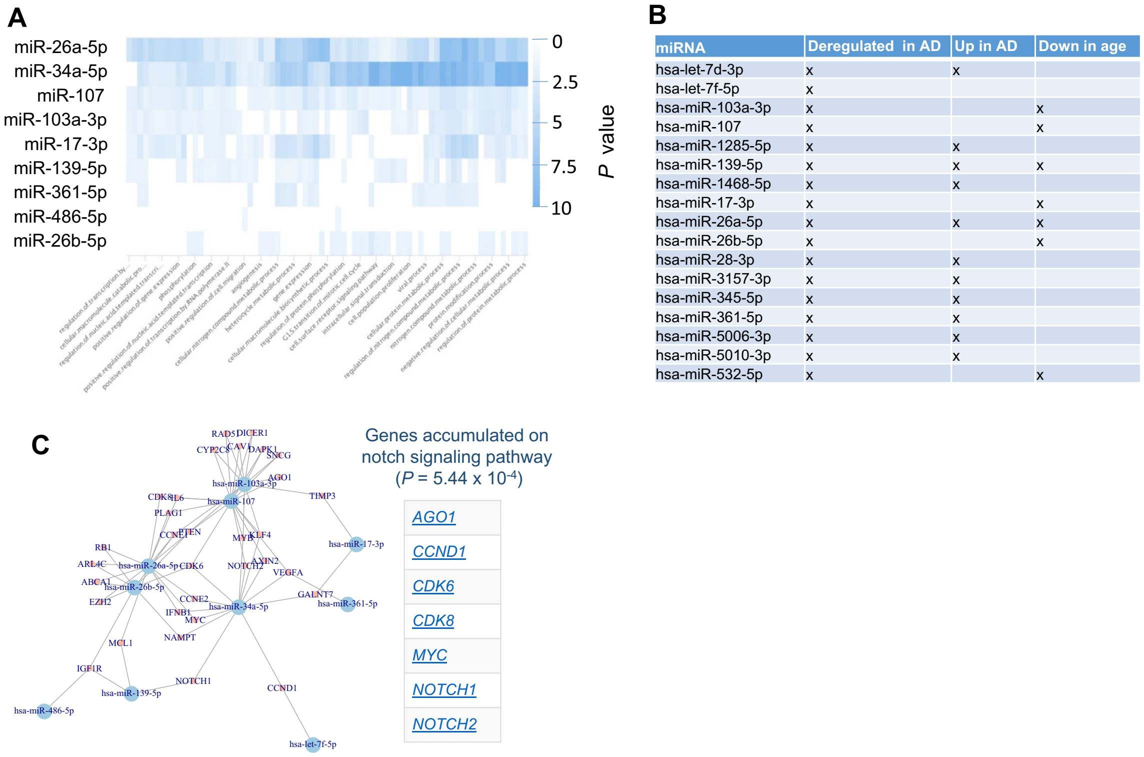


Figure 4 miRNA classifiers show a high diagnostic performance to detect AD
Diagnostic performance of the miRNA classifiers. **A.** ROC AUC for the diagnosis of AD patients compared to HC. **B.** ROC AUC for the diagnosis of AD patients compared to all controls combined (HC, MCI, and OND). The black line indicates the average ROC values of all replicates and folds of the 5 × 10-fold cross-validation models, and the gray area represents the resulting standard deviation. The average AUC obtained over all replicates and folds is displayed for each classification scenario. ROC, receiver operator characteristics; AUC, area under the curve.



miRNA	Deregulated in AD	Up in AD	Down in age
hsa-let-7d-3p	x	x	
hsa-let-7f-5p	x		
hsa-miR-103a-3p	x		x
hsa-miR-107	x		x
hsa-miR-1285-5p	x	x	
hsa-miR-139-5p	x	x	x
hsa-miR-1468-5p	x	x	
hsa-miR-17-3p	x		x
hsa-miR-26a-5p	x	x	x
hsa-miR-26b-5p	x		x
hsa-miR-28-3p	x	x	
hsa-miR-3157-3p	x	x	
hsa-miR-345-5p	x	x	
hsa-miR-361-5p	x	x	
hsa-miR-5006-3p	x	x	
hsa-miR-5010-3p	x	x	
hsa-miR-532-5p	x		x

Figure 5 AD miRNAs regulate distinct pathways and form a dense regulatory core network
A. Heatmap of the miRPathDB results. The heatmap presents the negative decade logarithm of miRNAs and target pathways, and the color represents the significance values. **B.** Overview of miRNAs in significant categories. For the three significant miEAA categories we highlight the miRNAs participating in the respective categories. **C.** miRNA target network from miRTargetLink. From miRTargetLink we extracted the target network of the miRNAs and generated a representation in R using the igraph library. Each node is a miRNA/gene and an edge means that the miRNA targets that gene. As an example of an enrichment of target genes, the genes on the Notch pathway are shown on the right side of the network.

miRNAs are enriched in specific functional categories

To get insights into the targeting of the dysregulated miRNAs, we performed different miRNA target analyses. First, we individually searched for each miRNA those pathways that are enriched with target genes of that miRNA. The result is presented as heat map in **Figure 5**A. Most significant pathways were computed for miR-34a-5p miR-26a-5p followed by miR-107. Among the pathways, many transcription regulated categories have been observed. This result is however to be expected since the main biological function of miRNAs is to regulate the gene expression.

To get more insights, we next performed a miRNA Enrichment analysis [22]. Following adjustment for multiple testing, we identified three categories to be significantly enriched including "Dys-regulation in AD" ($P = 4.8 \times 10^{-8}$), "Up-regulation in AD" ($P = 0.00018$), and "Age" ($P = 0.02$). Two of three categories were directly related to AD. Also this is an expected result for miRNAs that were known to be associated with AD. In addition, these miRNAs are negatively correlated with age. Although this was a weak correlation, it still suggests that the abundances of these miRNAs are lower in older patients. Figure 5B presents for each miRNA in the signature on which categories it has been observed. Performing an enrichment analysis for each of the three miRNAs clusters indicated in Figure 2B, we found cluster A to be especially enriched with miRNAs that are "up-regulated in AD" ($P = 4.9 \times 10^{-6}$) while for cluster B the only significant category was "down-regulated in AD" ($P = 0.04$).

In a third analysis we analyzed all target genes of the miRNAs that had strong evidence in the miRTarBase and were extracted from miRTargetLink. This analysis highlighted that for most miRNAs in our signature, target genes that have been experimentally validated are known. The target network shown in Figure 5C highlighted a dense structure. This network was enriched for genes associated with AD including ABCA1, DAPK1, IGF1R, and VEGFA according to the national institute of aging (NIA). Likewise, "DNA damage response" represented by CCND1, CCNE1, CCNE2, CDK6, MYC, RAD51, and RB1 was over represented. Moreover, the genes in that network were also enriched for the notch signaling pathway.

Discussion

In the current study we present results of our ongoing efforts to develop a diagnostic test for AD patients based on circulating miRNA profiles extracted from blood cells.

As Figure 1 and Table S1 highlight, the samples were largely homogenous with respect to the age and gender distribution. With respect to other characteristics the cohort was however heterogenous (*e.g.*, the origin of the samples from two continents, different diagnostic procedures to identify the patients, potentially different treatment regimens, or a spectrum of patients with higher and lower MMSE values). This heterogeneity helps us to understand whether the deregulation in miRNA patterns between AD patients and controls is of general nature and helps to assess whether *e.g.*, miRNAs are associated with the MMSE state.

The current outcomes are consistent with our previous studies in the US and Germany on smaller cohorts. In contrast to the previous studies relying on deep sequencing, we here applied RT-qPCR as molecular profiling technique that can be more easily driven towards application in clinical care. In the context of the known variability and the bias introduced by sample integrity and sample treatment [23–25] in deep sequencing data, RT-qPCR offers a promising alternative for routine application. But also for RT-qPCR experiments, there is a debate whether RNA samples with low integrity, *i.e.*, low RIN values, compromise miRNA expression data [26,27]. In our study, we also measured RIN as quality criterion for RNA integrity of the samples. The markers that we validated in this study seem to play partially an important role in different diseases. As an example, our most significant marker miR-532-5p is not only correlated and functionally associated to cancer [28–30]. The miRNA and its target network is also associated to sporadic amyotrophic lateral sclerosis [31]. Further, the miR-532-5p has also been discovered in exosomes of multiple sclerosis patients [32] and in exosomes of patients with the geriatric frailty syndrome [33]. Also, our analyses indicate a very essential role of exosome derived miRNAs.

The results of the two cohorts from the US and from Germany were highly concordant. As to be expected by the selection of AD-associated miRNAs for this study, the miRNAs and the target genes of the miRNAs were both significantly associated with the development of AD. Our test that is highly reproducible can be applied with a model based on specificity, sensitivity or trimmed for overall performance. The quality of the results is indicated by an AUC of 87.6% for the comparison between AD and unaffected controls, and an AUC of 83.5% for a comparison between AD and a combined group of unaffected controls, MCI patients and patients with OND. It is known that complex statistical learning approaches can lead to overfitting, especially considering the curse of dimensionality [34] and the fact that usually many more features (p) are measured as compared to the number of patients (n), the $p \gg n$ problem. In our study we however measured $p = 21$ markers and $n = 465$ individuals. Further, we even select small subsets of these markers for our models and perform comprehensive re-sampling to prevent potential overfitting. Although the de-regulation of miRNAs was generally concordant between the GER and the USA cohort, miRNAs have shown differences in the expression level in the two cohorts. This might be due to technical reasons such as shipment, other batch effects or biological differences. Despite this fact, the statistical learning approach succeeded to separate AD and controls in the GER and the USA cohort. In sum, the performance of our diagnostic solution compares well to other recently-developed tests, such as the plasma amyloid marker introduced by Nakamura and co-workers [4]. While already such single "omics" tests have a large potential, the targeted combination of few representatives from different "omics" classes can add even more diagnostic information, supporting clinicians in detecting AD patients in time. One challenge of respective studies is that the clinical diagnosis may be imperfect. The MCI patients that are an important second control group besides the unaffected controls may have already early forms of AD that are not yet detected with the current diagnostic means.

A pathway based analysis of miRNAs and target genes indicated a functional role of the miRNAs. This is further supported by a different blood compound distribution of those miRNAs that are up- and down-regulated in AD. Respective pathway analyses have however always considered with caution, especially when small sets of miRNAs are considered. Although the results of the analysis seem to be reasonable, a potential bias is hard to be excluded. *e.g.*, we picked already miRNAs known from literature to be associated with AD. An enrichment of AD related miRNAs itself is thus an expected result. Similarly, also the target gene analyses might be biased for miRNAs and target genes that are in the focus of many research groups.

As for other omics types, confounders including age and gender potentially influence also the results of miRNA biomarker studies [35]. To minimize the influence of such confounders, our cohorts largely show similar age and gender distribution (Table 1). In addition, we investigated the influence of the age and gender on the miRNA profiles. Except for a very modest influence of age, we found no evidence for an influence of these confounders on the miRNA pattern. Notably, miRNAs that are down-regulated in AD were partially expected to be lower expressed with increasing age in a normal population. Among the many different candidates for minimally-invasive and potentially early stage tests for AD, our study indicates that circulating miRNAs likely in combination with other blood-born omics profiles will contribute to stable tests applicable to specific diagnostic questions with regard to this highly complex disease.

Materials and methods

Overview of the study

In the current study we included patients from the US [8] and Germany [16] that were partially collected within the longitudinal Tübinger Erhebung von Risikofaktoren zur Erkennung von Neurodegeneration (TREND) study. From the former studies we included those individuals, where a sufficient amount of high-quality RNA was left for analysis. In detail, 169 individuals from our initial study (36%) [8], 107 individuals from the second study (23%) [16], as well as 189 newly collected individuals (41%) were included in the study. The studies were approved by the institutional review boards of Charité – Universitätsmedizin Berlin (EA1/182/10), or the ethical committee of the Medical Faculty of the University of Tuebingen (Nr. 90/2009BO2). All subjects gave written informed consent. Besides AD patients and HC, patients with OND such as Parkinson's disease (PD), schizophrenia or bipolar disorder were included and grouped together, termed OND. Further, patients with MCI were included to evaluate the specificity of the miRNA markers for AD. For each of the four groups and separately for the USA and GER cohorts, total number, age, gender distribution, and MMSE value are presented in Table 1. Moreover, from one individual, 12 technical replicates were measured continuously during the project as process control.

miRNA marker set selection

From our two previous studies [8,16] we selected the top miRNAs that were concordant between the two studies, and also checked for evidence that the miRNAs are associated with AD in literature. A final set of 21 miRNAs was selected. These are listed in Supplemental Table 4 where additional selection criteria are provided. In more detail, 17 miRNAs were significantly associated with AD in our first study, 14 miRNAs were significant in our second study. miR-34a-5p was not detected in our previous studies by NGS but in a study by Cosin-Tomas [36]. Further, this miRNA is one of our main targets regulating calcium signaling, NFKappaB pathway and T-cell killing and is down-regulated significantly in aging [37,38]. miR-151-3p is one of the most stable miRNAs in our studies as well as miR-486-5p, which is a red blood cell miRNA that serves as positive control [20].

RNA extraction and quality control

Total RNA from PAX-Gene Blood Tubes (Catalog No. 762165, BD Biosciences, Franklin Lakes, NJ) was isolated using the Qiacube robot with the PAXgene Blood miRNA Kit (Catalog No. 763134, Qiagen, Hilden, Germany) according to manufacturer's instructions. In the tubes, 2.5 ml blood are collected, typically yielding around 1 mg total RNA. RNA quantity and quality were assessed using Nanodrop (Thermo Fisher Scientific) and RNA Nano 6000 Bioanalyzer Kit (Catalog No. 5067-1511, Agilent Technologies, Santa Clara, CA). Mean RNA integrity number (RIN) value of the RNA samples was 7.5 (STDEV 1.4).

RT-qPCR

Quantification of miRNAs was performed using miScript PCR system and custom miRNA PCR arrays (all reagents from Qiagen, Hilden, Germany). Custom miRNA PCR arrays were designed in 96-well plates to measure the expression of 21 human miRNAs and RNU48 as well as RNU6 as two endogenous controls in duplicates. Two process controls (miR-TC for RT efficiency, PPC for PCR efficiency) were included as single probes. A total of 100 ng total RNA was used as input for reverse transcription reaction using miScriptRT-II kit according to manufacturer's recommendations in 20 µl total volume (Catalog No. 218161). Subsequently, 1 ng cDNA was used per PCR reaction. PCR reactions with a total volume of 20 µl were setup automatically using the miScript SYBR Green PCR system (Catalog No. 218076) in a Qiagility pipetting robot (Qiagen, Hilden, Germany) according to manufacturer's instructions. Data from samples that failed the quality criteria for the process controls was excluded, leaving expression data from 465 samples available for analysis. For process control over the course of the project, eleven technical replicates of one cDNA sample were measured throughout the course of the project to estimate technical reproducibility. We computed 55 pair-wise correlation coefficients between any pair of the replicates and found a median correlation of 0.996, indicating high technical reproducibility of our assay.

Statistical approaches

From the Cq values, delta Cq values in relation to the endogenous control (RNU48) were computed. Mean delta Cq value per individual was scaled to zero. Missing values were not imputed. As estimate of the expression on a linear scale, 2^{deltaCq} values were computed. For multi group comparisons, Analysis of Variance (ANOVA) was performed. Since the miRNA data and partially the response variable were not always normally distributed according to Shapiro Wilk tests, we performed for the pair-wise comparisons and for the correlation analysis parametric as well as non-parametric tests. For pair-wise comparisons, both, parametric t-test and non-parametric Wilcoxon Mann-Whitney test were calculated. If not mentioned explicitly and where applicable, all *P* values were adjusted for multiple testing by the Benjamini-Hochberg approach. For correlating miRNAs to the age and the MMSE value, the *P* value was computed based on parametric Pearson's product moment correlation coefficient as well as non-parametric Spearman Correlation. To find enrichment of miRNAs in specific blood compounds we used data of an NGS based blood cell miRNA repertoire [20]. Each miRNA was normalized to 100% and the different expression ratios in the different blood compounds were compared to each other.

miRNA target analysis

We performed three different approaches on miRNA target analysis. First, for each single miRNA the target pathways have been extracted from miRPathDB [39] and the CustomHeatmap tool was used to find miRNAs that target at least 5 pathways and pathways targeted by at least 5 miRNAs from biological GO processes. Next, we performed a so-called miRNA set enrichment analysis relying on the hypergeometric distribution using MIEAA [22]. Here, the miRNAs in the study were compared to the background distribution of all miRNAs and the procedure was repeated for the dysregulated miRNAs. All pathways with an adjusted *P* value below 0.05 were considered to be significant. Finally, we used the miRTargetLink tool [40] to extract the experimentally validated targets of the miRNAs. In this analysis only the strong target category from miRTarBase has been used to obtain specific results. From that data we computed a network using the R igraph package and performed an enrichment analysis of the target genes in that network.

Machine learning

A prediction model based on the RT-qPCR Cq values was developed using gradient boosted trees from the LightGBM framework (version 2.1.0). Since not all miRNAs were consistently measured for all patients, tree-based methods are particularly suited for this task, as they can handle missing values and no imputation is required. LightGBM ignores the missing values when computing the splits of the trees and assigns all samples with missing values to the side that reduces the loss most. The performance of the model was assessed using five repetitions of stratified ten-fold cross-validation using scikit-learn 0.19.1 with Python 3.6.4 [41]. Each repetition was initiated with an integer seed (0–4). Thus, in total 50 combinations of different training and validation sets were considered. The reported ROC AUC corresponds to the average performance over all repetitions and folds of the model, on data not used for training. The models were manually tuned (*i.e.*, no grid search was performed) over the number of leaves (testing ranges between 5 and 50), number of estimators (between 40 and 120), learning rate (0.01 to 0.2), and depth (3 to no restriction). The final model comparing patients with AD to all controls uses 30 leaves, a learning rate of 0.1 and 100 estimators. The model comparing patients with AD to unaffected controls uses 9 leaves, a learning rate of 0.05 and 100 estimators. The depth of both models was not restricted. Gradient boosted trees outperformed other tree-based methods such as random forests, or classifiers as Support Vector Machines or Neural Networks (data not shown). As an input for the classification task, the expression matrix of the delta Cq values has been used.

Data availability

The full data set is available as Table S1 without any restrictions.

Authors' contributions

NL measured the samples and supported the interpretation of data; TF and FK interpreted and analyzed the data; MG, WM, CS, AKvT, CD, FM, US, and DB supported the study conceptionally, added to the study protocol and enrolled patients; VK added to the clinical interpretation of the data; CB supported the data analysis and contributed to drafting the manuscript; SD and SG supported to measure and interpret the data; HPL, EM, AK were the PIs of the study, contributed to writing and correcting the manuscript and to data analysis and interpretation.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the Alzheimer Forschungs Initiative (AFI) (Grant No. AFI-Grant# 15013). This study was also supported by internal funds of Saarland University. We acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and Saarland University within the funding programme Open Access Publishing.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2019.09.004.

References

[1] Querfurth HW, LaFerla FM. Alzheimer's disease. N Engl J Med 2010;362:329–44.

[2] Weuve J, Hebert LE, Scherr PA, Evans DA. Deaths in the United States among persons with Alzheimer's disease (2010–2050). Alzheimers Dement 2014;10:e40–6.

[3] Murphy MP. Amyloid-Beta solubility in the treatment of Alzheimer's disease. N Engl J Med 2018;378:391–2.

[4] Nakamura A, Kaneko N, Villemagne VL, Kato T, Doecke J, Dore V, et al. High performance plasma amyloid-beta biomarkers for Alzheimer's disease. Nature 2018;554:249–54.

[5] Mapstone M, Cheema AK, Fiandaca MS, Zhong X, Mhyre TR, MacArthur LH, et al. Plasma phospholipids identify antecedent memory impairment in older adults. Nat Med 2014;20:415–8.

[6] Lunnon K, Sattlecker M, Furney SJ, Coppola G, Simmons A, Proitsi P, et al. A blood gene expression marker of early Alzheimer's disease. J Alzheimers Dis 2013;33:737–53.

[7] Fransquet PD, Lacaze P, Saffery R, McNeil J, Woods R, Ryan J. Blood DNA methylation as a potential biomarker of dementia: a systematic review. Alzheimers Dement 2018;14:81–103.

[8] Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol 2013;14:R78.

[9] Casanova R, Varma S, Simpson B, Kim M, An Y, Saldana S, et al. Blood metabolite markers of preclinical Alzheimer's disease in two longitudinally followed cohorts of older individuals. Alzheimers Dement 2016;12:815–22.

[10] Pichler S, Gu W, Hartl D, Gasparoni G, Leidinger P, Keller A, et al. The miRNome of Alzheimer's disease: consistent downregulation of the miR-132/212 cluster. Neurobiol Aging 2017;50: e1–10.

[11] Ren RJ, Zhang YF, Dammer EB, Zhou Y, Wang LL, Liu XH, et al. Peripheral blood microRNA expression profiles in Alzheimer's disease: screening, validation, association with clinical phenotype and implications for molecular mechanism. Mol Neurobiol 2016;53:5772–81.

[12] Denk J, Oberhauser F, Kornhuber J, Wiltfang J, Fassbender K, Schroeter ML, et al. Specific serum and CSF microRNA profiles distinguish sporadic behavioural variant of frontotemporal dementia compared with Alzheimer patients and cognitively healthy controls. PLoS One 2018;13:e0197329.

[13] Yang TT, Liu CG, Gao SC, Zhang Y, Wang PC. The serum exosome derived microRNA-135a, -193b, and -384 were potential Alzheimer's disease biomarkers. Biomed Environ Sci 2018;31:87–96.

[14] Hu YB, Li CB, Song N, Zou Y, Chen SD, Ren RJ, et al. Diagnostic value of microRNA for Alzheimer's disease: a systematic review and meta-analysis. Front Aging Neurosci 2016;8:13.

[15] Nagaraj S, Zoltowska KM, Laskowska-Kaszub K, Wojda U. microRNA diagnostic panel for Alzheimer's disease and epigenetic trade-off between neurodegeneration and cancer. Ageing Res Rev 2019;49:125–43.

[16] Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, et al. Validating Alzheimer's disease micro RNAs using next-generation sequencing. Alzheimers Dement 2016;12:565–76.

[17] Leidinger P, Brefort T, Backes C, Krapp M, Galata V, Beier M, et al. High-throughput qRT-PCR validation of blood microRNAs in non-small cell lung cancer. Oncotarget 2016;7:4611–23.

[18] Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, et al. Distribution of miRNA expression across human tissues. Nucleic Acids Res 2016;44:3865–77.

[19] Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. Nat Methods 2011;8:841–3.

[20] Juzenas S, Venkatesh G, Hubenthal M, Hoeppner MP, Du ZG, Paulsen M, et al. A comprehensive, cell specific microRNA catalogue of human peripheral blood. Nucleic Acids Res 2017;45:9290–301.

[21] Swarbrick S, Wragg N, Ghosh S, Stolzing A. Systematic review of miRNA as biomarkers in Alzheimer's disease. Mol Neurobiol 2019;56:6156–67.

[22] Backes C, Khaleeq QT, Meese E, Keller A. miEAA: microRNA enrichment analysis and annotation. Nucleic Acids Res 2016;44:W110–6.

[23] Backes C, Sedaghat-Hamedani F, Frese K, Hart M, Ludwig N, Meder B, et al. Bias in high-throughput analysis of miRNAs and implications for biomarker studies. Anal Chem 2016;88:2088–95.

[24] Ludwig N, Becker M, Schumann T, Speer T, Fehlmann T, Keller A, et al. Bias in recent miRBase annotations potentially associated with RNA quality issues. Sci Rep 2017;7:5162.

[25] Backes C, Leidinger P, Altmann G, Wuerstle M, Meder B, Galata V, et al. Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. Anal Chem 2015;87:8910–6.

[26] Becker C, Hammerle-Fickinger A, Riedmaier I, Pfaffl MW. mRNA and microRNA quality control for RT-qPCR analysis. Methods 2010;50:237–43.

[27] Jung M, Schaefer A, Steiner I, Kempkensteffen C, Stephan C, Erbersdobler A, et al. Robust microRNA stability in degraded RNA preparations from human tissue and cell samples. Clin Chem 2010;56:998–1006.

[28] Yamada Y, Arai T, Kato M, Kojima S, Sakamoto S, Komiya A, et al. Role of pre-miR-532 (miR-532-5p and miR-532-3p) in regulation of gene expression and molecular pathogenesis in renal cell carcinoma. Am J Clin Exp Urol 2019;7:11–30.

[29] Xie X, Pan J, Han X, Chen W. Downregulation of microRNA-532-5p promotes the proliferation and invasion of bladder cancer cells through promotion of *HMGB3*/Wnt/beta-catenin signaling. Chem Biol Interact 2019;300:73–81.

[30] Wei H, Tang QL, Zhang K, Sun JJ, Ding RF. miR-532-5p is a prognostic marker and suppresses cells proliferation and invasion by targeting *TWIST1* in epithelial ovarian cancer. Eur Rev Med Pharmacol Sci 2018;22:5842–50.

[31] Liguori M, Nuzziello N, Introna A, Consiglio A, Licciulli F, D'Errico E, et al. Dysregulation of microRNAs and target genes networks in peripheral nlood of patients with sporadic Amyotrophic Lateral Sclerosis. Front Mol Neurosci 2018;11:288.

[32] Selmaj I, Cichalewska M, Namiecinska M, Galazka G, Horzelski W, Selmaj KW, et al. Global exosome transcriptome profiling reveals biomarkers for multiple sclerosis. Ann Neurol 2017;81:703–17.

[33] Ipson BR, Fletcher MB, Espinoza SE, Fisher AL. Identifying exosome-derived microRNAs as candidate biomarkers of Frailty. J Frailty Aging 2018;7:100–3.

[34] Barbour DL. Precision medicine and the cursed dimensions. NPJ Digit Med 2019;2:4.

[35] Meder B, Backes C, Haas J, Leidinger P, Stahler C, Grossmann T, et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. Clin Chem 2014;60:1200–8.

[36] Cosin-Tomas M, Antonell A, Llado A, Alcolea D, Fortea J, Ezquerra M, et al. Plasma miR-34a-5p and miR-545-3p as early biomarkers of alzheimer's disease: potential and limitations. Mol Neurobiol 2017;54:5550–62.

[37] Hart M, Walch-Ruckheim B, Krammes L, Kehl T, Rheinheimer S, Tanzer T, et al. miR-34a as hub of T cell regulation networks. J Immunother Cancer 2019;7:187.

[38] Hart M, Walch-Ruckheim B, Friedmann KS, Rheinheimer S, Tanzer T, Glombitza B, et al. miR-34a: a new player in the regulation of T cell function by modulation of NF-kappaB signaling. Cell Death Dis 2019;10:46.

[39] Backes C, Kehl T, Stockel D, Fehlmann T, Schneider L, Meese E, et al. miRPathDB: a new dictionary on microRNAs and target pathways. Nucleic Acids Res 2017;45:D90–6.

[40] Hamberg M, Backes C, Fehlmann T, Hart M, Meder B, Meese E, et al. MiRTargetLink–miRNAs, genes and interaction networks. Int J Mol Sci 2016;17:564.

[41] Pedregosa F, Varoquaux G, Gramfort A, Michel V, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

3.24 *Deep sequencing of small non-coding RNAs reveals hallmarks and regulatory modules of the transcriptome during Parkinson's disease progression*

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

This article is available under: https://doi.org/10.1038/s43587-021-00042-6

3.25 *Evaluating the use of circulating microRNA profiles for lung cancer detection in symptomatic patients*

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

This article is available under: https://doi.org/10.1001/jamaoncol.2020.0001

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19665-1 OPEN

Common diseases alter the physiological age-related blood microRNA profile

Tobias Fehlmann[1], Benoit Lehallier[2], Nicholas Schaum[2], Oliver Hahn[2], Mustafa Kahraman[1], Yongping Li[1], Nadja Grammes[1], Lars Geffers[3], Christina Backes[1], Rudi Balling[3,4,5], Fabian Kern[1], Rejko Krüger[3,4,5], Frank Lammert[6], Nicole Ludwig[7], Benjamin Meder[8], Bastian Fromm[9], Walter Maetzler[10], Daniela Berg[10], Kathrin Brockmann[11], Christian Deuschle[11], Anna-Katharina von Thaler[11], Gerhard W. Eschweiler[12], Sofiya Milman[13], Nir Barziliai[13], Matthias Reichert[6], Tony Wyss-Coray[2], Eckart Meese[7] & Andreas Keller[1,2,14]✉

Aging is a key risk factor for chronic diseases of the elderly. MicroRNAs regulate post-transcriptional gene silencing through base-pair binding on their target mRNAs. We identified nonlinear changes in age-related microRNAs by analyzing whole blood from 1334 healthy individuals. We observed a larger influence of the age as compared to the sex and provide evidence for a shift to the 5′ mature form of miRNAs in healthy aging. The addition of 3059 diseased patients uncovered pan-disease and disease-specific alterations in aging profiles. Disease biomarker sets for all diseases were different between young and old patients. Computational deconvolution of whole-blood miRNAs into blood cell types suggests that cell intrinsic gene expression changes may impart greater significance than cell abundance changes to the whole blood miRNA profile. Altogether, these data provide a foundation for understanding the relationship between healthy aging and disease, and for the development of age-specific disease biomarkers.

[1] Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany. [2] Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA. [3] Luxembourg Center for Systems Biomedicine, 4362 Esch-sur-Alzette, Luxemburg. [4] Transversal Translational Medicine, Luxembourg Institute of Health (LIH), 1445 Strassen, Luxembourg. [5] Parkinson Research Clinic, Centre Hospitalier de Luxembourg, 1210 Luxembourg, Luxemburg. [6] Internal Medicine, Saarland University, 66421 Homburg, Germany. [7] Human Genetics, Saarland University, 66421 Homburg, Germany. [8] Internal Medicine, University Hospital Heidelberg, 69120 Heidelberg, Germany. [9] Department of Molecular Biosciences, Stockholm University, 11418 Stockholm, Sweden. [10] Department of Neurology, Christian-Albrechts-Universität zu Kiel, 24105 Kiel, Germany. [11] TREND study center Tübingen, Tübingen, Germany. [12] Geriatric Center and the Department of Psychiatry and Psychotherapy, University Hospital Tübingen, 72076 Tübingen, Germany. [13] The Institute for Aging Research, Albert Einstein College of Medicine, New York, NY 10461, USA. [14] Center for Bioinformatics, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany. ✉email: andreas.keller@ccb.uni-saarland.de

Aging is the leading risk factor for cardiovascular disease, diabetes, dementias including Alzheimer's disease, and cancer, together accounting for the majority of debilitating illnesses worldwide[1]. Uncovering common therapeutic targets to prevent or treat these diseases simultaneously could convey enormous benefits to quality of life. It is therefore essential to model the cellular processes culminating in these diverse maladies through an understanding of the molecular changes underlying healthy and pathological aging[2]. Accordingly, a variety of molecular studies have been conducted in humans, including whole genome analysis of long-lived individuals[3], transcriptomic analyses of tissues[4], plasma proteomic profiling[5], and the exploration of epigenetic control of aging clocks[6]. Recent organism-wide RNA-sequencing data of whole organs and single cells across the mouse lifespan provide an important and complementary database from which to build models of molecular cascades in aging[7,8].

Functional improvement of aged tissues has been achieved by an expanding number of techniques, ranging from dietary restriction[9] to senescent cell elimination and partial cellular reprogramming. This also includes heterochronic parabiosis, in which an old mouse is exposed to a young circulatory system. These experiments point to systemic factors in the blood of young mice that modulate organ function in aged animals[10,11]. Indeed, the list of individual plasma proteins with beneficial or detrimental effects on different tissues is growing. It is likely, however, that each plasma protein interacts with complex intracellular regulatory networks, and that alterations to such networks are a key component of aging and rejuvenation.

Non-coding ribonucleic acids like microRNAs (miRNAs) represent essential players governing these molecular cascades, and they show a highly complex spectrum of biological actions[12–14]. MicroRNAs are a family of short single stranded non-coding RNA molecules that regulate post-transcriptional gene silencing through base-pair binding on their target mRNAs[13], thereby regulating most if not all cellular and biological processes[15]. Yet, their involvement in the aging process and rejuvenation of aged tissues is often ignored by transcriptomic studies and is thus largely uncharacterized. A single microRNA targets not only untranslated regions (UTRs) of numerous genes, but it can also bind multiple sites within a single UTR[16]. Similarly, a UTR of a specific gene can contain target sites for dozens or even hundreds of miRNAs. Since their discovery, miRNA changes have been reported for almost all cancers and many non-cancer diseases like Alzheimer's disease[17,18], multiple sclerosis[19], or heart failure[20]. And although relatively sparse, several studies have measured aging miRNA expression in different human and primate tissues[21]. For example, Somel and co-workers analyzed miRNA, mRNA, and protein expression linked to development and aging in the prefrontal cortex of humans and rhesus macaques over the lifespan[22]. Likewise, changes of miRNA levels in aging human skeletal muscle have been characterized[23], as have miRNA levels in body fluids such as serum[24,25]. In whole blood, we previously reported a significant number of age-related miRNAs[26], and Huan and co-workers measured a selection of miRNAs by RT-qPCR in whole blood from over 5000 individuals from the Framingham Heart Study[27]. While these initial studies are intriguing, they can be limited by the use of discrete time points, incomplete lifespan coverage, limited cohort sizes, and incomplete miRNA panels.

Here, we performed a comprehensive characterization of all 2549 annotated miRNAs (miRBase V21) in 4393 whole blood samples from both sexes across the lifespan (30–90 years). To understand the relationship between healthy aging and disease, we included 1334 healthy controls (HC), 944 patients with Parkinson's disease (PD), 607 with heart diseases (HD), 586 with non-tumor lung diseases (NTLD), 517 with lung cancer (LC), and 405 with other diseases (OD) (Fig. 1a, b; Supplementary Data 1).

Results

miRNA profiles are stronger associated with the age as compared to the sex. We first sought to model healthy aging as a baseline for understanding disease. As males have shorter lifespans than females, and each sex suffers a different array of age-related diseases, we investigated the interplay between age and sex on blood miRNA profiles. Confirming our previous observation in a cohort of 109 individuals[26], we found that age has a more pronounced influence than sex. In fact, 1568 miRNAs significantly correlated with age, but only 362 correlated with sex according to Benjamini–Hochberg adjusted p-values of the Wilcoxon Mann–Whitney test (Fig. 2a, b). While 231 miRNAs overlapped between these groups, this number was not significant (two-sided Fisher's exact test p-value of 0.35; Pearson's Chi-squared Test of 0.36), suggesting that, in general, those miRNAs changing with age are shared by both sexes, and those specific to one sex do not change with age. In consequence, the Spearman correlation coefficient (SC) of age-related changes between males and females was high (SC of 0.884, $p < 10^{-16}$, Fig. 2c).

We next sorted miRNAs by their correlation with age, regardless of their significance, and assigned each to one of 5 groups: strongly decreasing with age (cluster 1: 174 miRNAs, SC < -0.2), moderately decreasing (cluster 2: 382 miRNAs; $-0.2 < \mathrm{SC} < -0.1$), unaltered (cluster 3: 1451 miRNAs; $-0.1 < \mathrm{SC} < 0.1$), moderately increasing (cluster 4: 368 miRNAs; $0.1 < \mathrm{SC} < 0.2$), and strongly increasing (cluster 5: 174 miRNAs, $\mathrm{SC} > 0.2$) (Supplementary Data 2). As miRNAs regulate a diverse array of critical pathways[28], we performed microRNA enrichment analysis and annotation (miEAA) on this sorted list, thereby calculating a running sum of miRNAs associated with each of ~14,000 biochemical categories and pathways. We revealed a remarkable disequilibrium between the number of pathways related to downregulated miRNAs (76 pathways) and upregulated miRNAs (620 pathways; adjusted p-value < 0.05; Supplementary Data 3). This is even more striking considering the number of miRNAs increasing or decreasing did not differ significantly (556 with $\mathrm{SC} < -0.1$; 542 with $\mathrm{SC} > 0.1$), and suggests that miRNAs increasing with age have a higher functional relevance. Reassuringly, for miRNAs decreasing with age we found "Negative Correlated with Age" ($p = 4 \times 10^{-10}$) among the most significant categories (Fig. 2d). A large fraction of the top pathways regardless of the miRNA direction were enriched for brain function and neurodegeneration, including "Downregulated in Alzheimer's Disease" ($p = 10^{-5}$), "regulation of synaptic transmission" ($p = 0.028$), and "APP catabolic processes" ($p = 0.032$) (Fig. 2e, Supplementary Fig. 1a–l).

Although such linear correlation analyses can reveal meaningful biological features, the importance of nonlinear aging changes, such as those found for plasma proteins[5] and tissue gene expression, is becoming increasingly evident. We therefore aimed to use the high temporal resolution of the dataset to more thoroughly understand whole blood miRNA dynamics across the lifespan. We first plotted miRNA trajectories for each of the 5 clusters (Supplementary Fig. 2), confirming many miRNAs exhibit non-linear patterns. By comparing linear and nonlinear correlations for each, we uncovered nonlinear changes in 116 of the 1098 miRNAs altered with age, of which 90 decreased and 26 increased (Fig. 2f, g, Supplementary Data 4). A miEAA analysis highlighted a significant enrichment of miRNAs following nonlinear trajectories with aging in basically all human tissues[29] (Fig. 2h). This finding stands out considering the high degree of tissue specificity of miRNAs. We thus speculate that diseases

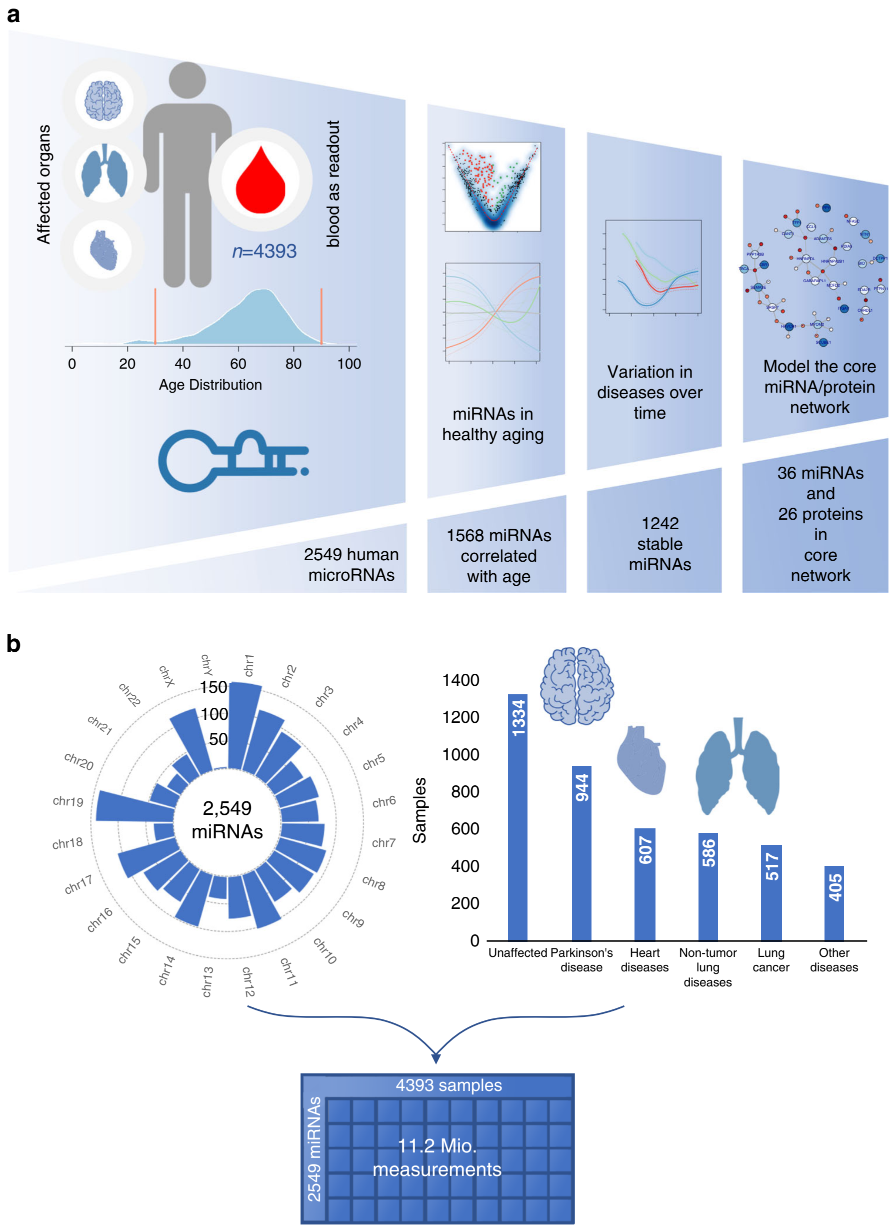
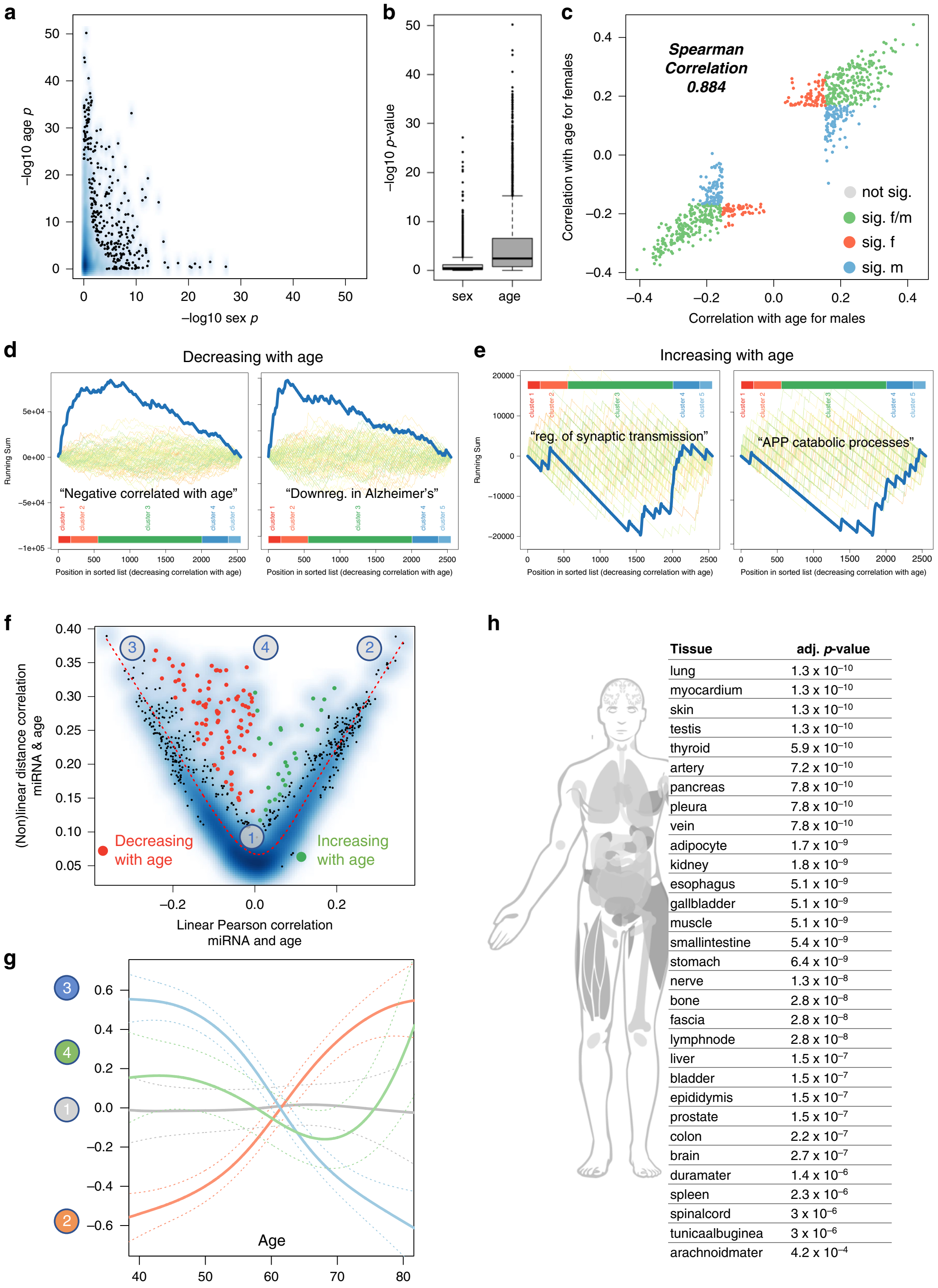


Fig. 1 Study characteristics. a Study set up and analysis workflow from high-throughput data to a specific aging network. The cohort consist of 4393 samples of which the age distribution is provided. For the 4393 samples genome wide miRNA screening using microarrays has been performed. The first analysis describes 1568 miRNAs that are correlated to age in healthy individuals. In the second step we identified disease specific miRNA changes with aging and finally define a set of 1242 miRNAs that are not affected by diseases. Finally, to model regulatory cascades in healthy aging we related the miRNA data to plasma proteins and identified a core aging network. **b** The circular plot shows the genome wide nature of our miRNA approach, all miRNAs from miRBase V21 were included in the experimental analysis. We measured 4393 samples for the abundance of these miRNAs, resulting in a 2549 times 4393 data table containing 11.2 million miRNA measurements that correspond to over 2 × 10^8 spots on the arrays.

affecting these organs might be associated with changes in blood miRNA profiles.

miRNA arm shifts are associated with aging. A shift in the expression of the 3' and 5' mature arm of miRNAs is observed between different tissues[30] tissues but also in healthy and diseased conditions such as cancer[31]. We speculated that likewise aging may affect the arm distribution and searched for respective arm shift events. Indeed, we observed a correlation of the arm specific expression in 40 cases (Supplementary Data 5). For 27 miRNAs (67.5%) we observed increasing 5' mature expression and decreasing 3' expression over age while in 13 cases 32.5% of cases the 3' form increased and the 5' form decreased. These results indicate a generally increasing 5' mature miRNA expression with aging. The largest absolute increase of 5' mature expression was identified for miR-6786. A miRSwitch analysis highlighted that usually the 3' form is dominating in H. sapiens



Tissue	adj. *p*-value
lung	1.3 x 10^{-10}
myocardium	1.3 x 10^{-10}
skin	1.3 x 10^{-10}
testis	1.3 x 10^{-10}
thyroid	5.9 x 10^{-10}
artery	7.2 x 10^{-10}
pancreas	7.8 x 10^{-10}
pleura	7.8 x 10^{-10}
vein	7.8 x 10^{-10}
adipocyte	1.7 x 10^{-9}
kidney	1.8 x 10^{-9}
esophagus	5.1 x 10^{-9}
gallbladder	5.1 x 10^{-9}
muscle	5.1 x 10^{-9}
smallintestine	5.4 x 10^{-9}
stomach	6.4 x 10^{-9}
nerve	1.3 x 10^{-8}
bone	2.8 x 10^{-8}
fascia	2.8 x 10^{-8}
lymphnode	2.8 x 10^{-8}
liver	1.5 x 10^{-7}
bladder	1.5 x 10^{-7}
epididymis	1.5 x 10^{-7}
prostate	1.5 x 10^{-7}
colon	2.2 x 10^{-7}
brain	2.7 x 10^{-7}
duramater	1.4 x 10^{-6}
spleen	2.3 x 10^{-6}
spinalcord	3 x 10^{-6}
tunicaalbuginea	3 x 10^{-6}
arachnoidmater	4.2 x 10^{-4}

with 5' dominance mostly in plasma samples. For the miRNA with the most decreasing 5' expression ratio (miR-4423) we found dominating 3' expression mostly in breast milk, the heart, testis, stem cells and blood cells. Our results thus suggest an altered ratio of the 3' to 5' mature expression ratio that might be attributed to or effect different tissues.

The association between age and miRNA expression is partially lost in diseases. Although the cellular and molecular degeneration of aging often instigates age-related disease, there are nonetheless elderly individuals who have lived entirely disease-free lives. We therefore asked what differentiates such healthy aging from aging resulting in disease. For each disease and healthy controls, we

Fig. 2 miRNAs dependency on age and gender. a Smoothed scatter plot of the two-tailed age and gender association *p*-value for 2549 miRNAs. *P*-values for the sex are computed using Wilcoxon Mann–Whitney test and for the Spearman Correlation via the asymptotic t approximation. The *p*-values are Benjamini–Hochberg adjusted. **b** Boxplot of the age and gender *p*-value from **a** for 2549 miRNAs. The box spans the 25% and 75% quantile, the solid horizontal line represents the median and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. **c** Correlation of miRNAs with age in males and females. Gray dots: not significant; orange and blue dots: miRNAs significantly correlated with age only in males or females; green dots: miRNAs significantly correlated with age in males and females. **d** Results of the miRNA enrichment analysis. Colored curves in the background represent random permutations of miRNAs. The cluster membership is projected next to the order of miRNAs. The category "negative correlated with age" is highly significant and confirms our data in general. Also, the category "downregulated in AD" is enriched with miRNAs decreasing over age. **e** Regulation of synaptic transmission is among the categories being enriched in miRNAs going up with age. Moreover, APP catabolic processes is another category being enriched in miRNAs going up with age. **f** Linear Pearson correlation versus non-linear distance correlation for the association of age to miRNAs. Orange dots have a high non-linear correlation that is not explained by linear correlation and are decreasing with age, green dots have a high non-linear correlation that is not explained by linear correlation and are increasing with. The orange dotted line represents a smoothed spline and the four numbers in gray circles represent the position of miRNAs where examples are provided in **g**. **g** Examples of correlation for miRNAs with age. (1) gray: no correlation; (2) orange dominantly positive linear correlation; (3) blue dominantly negative linear correlation; (4) non-linear correlation. Each solid line is a smoothing spline. **h** Tissue enrichment for the miRNAs that are correlated with age in a non-linear fashion. The human model has all organs highlighted in gray that are significantly enriched. The table on the right lists the organs with corresponding *p*-values. *P*-values have been computed using the hypergeometric distribution and were adjusted for multiple testing using the Benjamini–Hochberg approach.

computed the Spearman correlation (SC) with age for all 2549 miRNAs (Fig. 3a, Supplementary Data 6). Overall, healthy controls reached the largest absolute SC, greater than twice that of the pooled disease cohort, and larger than any individual disease. Using an Analysis of variance, we found highly significant differences ($p < 2.2 \times 10^{-16}$) and a non-parametric Wilcoxon Mann–Whitney test confirmed the significant differences of absolute Spearman correlation in healthy versus diseased samples ($p < 2.2 \times 10^{-16}$). In line with these findings, samples from healthy individuals showed far more miRNAs with significant age correlations (Fig. 3b), suggesting that the presence of an age-related disease may disrupt healthy aging miRNA profiles (Wilcoxon Mann–Whitney test $p < 2.2 \times 10^{-16}$). For example, lung cancer patients were enriched for a positive correlation with age, while miRNAs in patients with heart disease were enriched for negative correlation with age. We then compared the miRNA trajectories from the 5 clusters of healthy individuals to the matched clusters in diseased patients (Supplementary Fig. 2), and similarly, miRNAs from diseased individuals show far weaker aging patterns. This held true both when each disease was analyzed separately, or pooled.

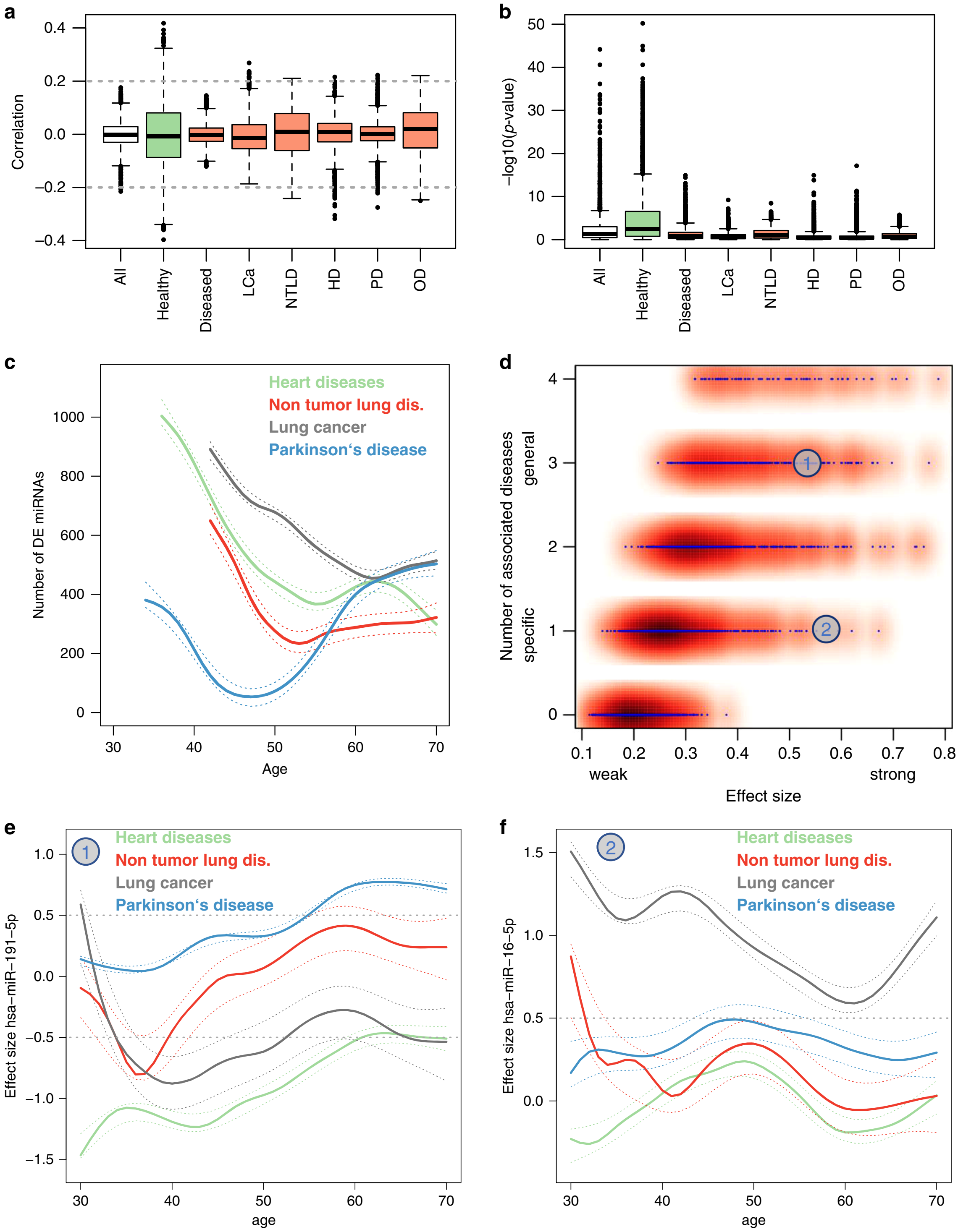
To determine the extent to which diseases affect miRNA abundance compared to healthy controls, we computed the number of differentially expressed miRNAs between cases and controls using a sliding window analysis. That is, we first compared diseased individuals aged 30–39 years to healthy individuals aged 30–39 years, then increased the window in increments of one year (31–40 years, 32–41 years, etc.) to the final window of 70–79 years (Fig. 3c, Supplementary Fig. 3a, b). As the age distribution varied between these groups, we excluded any window in which there were fewer than 20 disease cases and 20 healthy controls. Interestingly, for all diseases the number of differentially expressed miRNAs was high in young adults but decreased sharply into middle age, plateauing around age 60 for lung cancer and 50 for non-tumor lung diseases. Heart diseases largely plateaued by the early 50s. Parkinson's disease (PD), on the other hand, reached a minimum around age 47 before sharply increasing. With the exception of PD, these data show that aged healthy and diseased individuals are more similar than younger healthy and diseased individuals, perhaps suggesting that aged healthy individuals share some phenotypic characteristics of heart and lung disease.

We next asked if these diseases shared any miRNA alterations, and surprisingly we found that those miRNAs most commonly dysregulated were also those with the largest effect size (Fig. 3d). These pan-disease miRNAs included miR-191-5p (Fig. 3e), which targets mRNAs involved in cellular senescence[28]. We also observed disease-specific miRNAs like miR-16-5p, which targets the PI3K-Akt signaling pathway and microRNAs involved in lung cancer[28]. In summary, miRNA expression seems to be orchestrated in healthy aging with a loss of regulation in disease. In addition to disease-specific miRNAs, there appears to be a group of pan-disease miRNAs that change in a distinct manner. We thus asked on the specificity of biomarkers for diseases, especially in an age dependent context.

Distinct miRNA biomarker sets exist in young and old patients. The previous analyses of biomarkers in diseases were largely quantitative, i.e., we computed the number of dysregulated miRNAs in diseases for young and old patients. Here, we set to evaluate changes in the miRNA sets for young and old patients in the diseases. In this context we made use of the dimension reduction and visualization capabilities of self-organizing maps (SOMs). First, we considered the effect sizes of miRNAs for the two most global comparisons, i.e., healthy controls versus diseases and old (60–79 years) versus young (30–59 years) individuals. The heat map representation for the healthy versus disease comparison (Fig. 4a) and for young versus old individuals (Fig. 4b) highlights distinct patterns for the two comparisons and indicates that the aging miRNAs are different from the general disease miRNAs. This analysis however calls for a disease specific consideration. To this end we computed for each of the four diseases biomarkers in old and young patients using again the effect size as performance indicator and the self-organizing map analysis followed by a hierarchical clustering (Fig. 4c). While the cluster heat maps identify larger differences between the disease biomarker sets as compared to young and old biomarkers, also the sets within the diseases vary greatly (Fig. 4c). In line with the previous analyses we observe larger effects for all diseases but PD in young patients (middle row of Fig. 4c). In old patients, the respective biomarkers are partially lost. Only in few cases new biomarkers emerge in old patients that are not present in young patients. As the full annotation of the SOM grid shows, each SOM cell has an average of 8 cluster members with a standard deviation of 3.5 miRNAs (Supplementary Data 7). The distribution largely corresponds to a normal distribution, only four cells (24, 62, 81, and 82 in Supplementary Data 7) contain more than 15 miRNAs (mean + two times the standard deviation).

The previous analyses suggest distinct biomarker sets for young and old patients in the different diseases. As a consequence, future biomarker test based on miRNAs may not only be established for a disease but for a specific age range of patients with that disease.

Given the results from this and the previous section we computed for each miRNA in each disease and each age window



the effect size (Supplementary Data 8). The respective supplementary data provides detailed insights in how specific certain miRNAs are for specific diseases and age ranges and can support ongoing biomarker studies significantly.

All results obtained so far argue for a strong immunological component of the miRNAs, and as a consequence of miRNA target networks. Since our experimental system profiles whole blood miRNAs, we set out to determine the cellular origin by computational deconvolution.

White blood cells are the major repository of miRNAs in whole blood. Circulating immune cells have been implicated in aging and a variety of age-related diseases, and one of the most

Fig. 3 Diseases miRNAs are affected by age effects. a Boxplot of the Spearman correlation coefficient for each miRNA to all samples, healthy individuals, and patients. Group sizes: $n_{HC} = 1334$, $n_{PD} = 944$, $n_{HD} = 607$, n_{NTLD}, $n_{LC} = 517$, $n_{OD} = 405$. The box spans the 25% and 75% quantile, the solid horizontal line represents the median and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. **b** Boxplot of *p*-values for the Spearman correlation coefficient of each miRNA to all samples, healthy individuals, and patients from **a**. Group sizes: $n_{HC} =$ 1334, $n_{PD} = 944$, $n_{HD} = 607$, n_{NTLD}, $n_{LC} = 517$, $n_{OD} = 405$. The box spans the 25% and 75% quantile, the solid horizontal line represents the median and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The *p*-values have been computed via the asymptotic t approximation. **c** Number of deregulated miRNAs in disease groups depending on different ages in a sliding window analysis. Each solid line is a smoothing spline (green–heart diseases; red–non tumor lung diseases; gray–lung cancer; blue–Parkinson's disease). The areas represent the 95% confidence intervals. For all disease groups, the number of deregulated miRNAs decreases with age while it increases for Parkinson's Disease. **d** Smoothed scatterplot showing the average effect size per miRNA dependent on the number of diseases where the miRNA is associated with. In the lower right corner (the *y*-axis value of 1) the specific miRNAs with high effect sizes can be found. In the upper right corner, miRNAs with high effect sizes independent of the disease are located. The two numbers represent the location of the examples provided in **e** and **f**. **e** Example of a miRNA that is downregulated in heart diseases of younger patients, upregulated in older Parkinson's patients and not deregulated in lung diseases. Each solid line is a smoothing spline (green–heart diseases; red–non tumor lung diseases; gray–lung cancer; blue–Parkinson's disease). The areas represent the 95% confidence intervals. **f** Example of a miRNA from the lower right part of Fig. 3d. The miRNA is significant upregulated in lung cancer independent of age but basically not associated with other diseases. Color codes of panels **c**, **e**, and **f** are matched. Each solid line is a smoothing spline (green–heart diseases; red–non tumor lung diseases; gray–lung cancer; blue–Parkinson's disease). The areas represent the 95% confidence intervals.

common diagnostic tests for disease is blood cell profiling. Since miRNAs are known to be enriched in different blood cell types[32], we performed computational deconvolution of the whole blood miRNA profile, thereby grouping miRNAs by their predicted cell type(s) of origin (Fig. 5a). A total of 196 miRNAs were attributed to one specific cell type, including 127 miRNAs arising from monocytes. Most others derive from three or more types. For example, the largest group of 139 miRNAs stems from a combination of white and red blood cells (WBCs, RBCs), exosomes, and serum. And the third largest group of 119 is restricted to six types of WBCs. We also observed 31 miRNAs specific for NK cells, 19 specific for T-helper cells, 11 specific for B cells, and 8 specific for cytotoxic T cells. Overall, for those miRNAs for which we could assign a prospective origin, we found WBCs as the main contributor, even though they represent a substantially smaller volume of whole blood relative to RBCs and serum (Fig. 5b).

We then applied this analysis to those miRNAs changing with age, and found that those increasing appear to largely originate from B cells, monocytes, NK cells, cytotoxic T cells, and serum (Fig. 5c). In contrast, miRNAs decreasing with age are those enriched in neutrophils, T helper cells, and RBCs. These data indicate shifts in aging miRNA trajectories of specific blood cell types (Supplementary Fig. 4). Interestingly, for the above cell types, known age-related abundance changes largely follow opposite trends: lymphocytes generally decrease with age while neutrophils increase with age[33]. This suggests that cell-intrinsic gene expression changes age may significantly contribute to the observed whole blood miRNA profiles.

miRNAs associated with healthy aging regulate the expression of plasma proteins. An increasing body of evidence points to functional roles of systemic plasma proteins in aging and disease[5]. These proteins may represent downstream targets of blood-borne miRNAs. We thus compared our data to a recent dataset of plasma proteins associated with age in healthy individuals[5]. Because miRNAs regulate genes/proteins in a complex network, miRNAs increasing with age do not necessarily lead to downregulation of all target genes/proteins, and vice versa. Accordingly, we observed only one tendency: miRNAs decreasing with age (cluster 1 and 2) showed a slight enrichment for regulating proteins increasing with age (Fig. 6a). Considering such complexity, we employed a network-based analysis. Using all pairwise interactions of miRNAs with plasma proteins, we first computed a regulatory network (Fig. 6b). From this, we extracted a core network containing the top 5% downregulated miRNAs and the top 5% upregulated proteins, which was then further refined by including only experimentally validated miRNA/target genes mined from the literature[34], as well as miRNA/target pairs with an absolute Spearman correlation of at least 0.6. This stringent core network consists of 36 miRNAs targeting 26 genes (proteins) and splits into two larger and six smaller connected components (Fig. 6c). The densest part of the core network contains the axon guidance related semaphorin 3E (SEMA3E) and serine and arginine rich splicing factor 7 (SRSF7), which were targeted by 8 miRNAs including miR-6812-3p (Fig. 6d, Supplementary Fig. 5, Supplementary Fig. 6). Intriguingly, there exist no studies of this miRNA, but it targets SEMA3E in an age dependent manner with a Spearman correlation of −0.89.

Finally, we investigated the possible cell type of origin of these core miRNAs with deconvolution, which showed enrichment for neutrophils, monocytes, and B cells (Fig. 6e). We then used single-cell PBMC transcriptomic data to determine if SEMA3A or SRSF7 were expressed in these same cell types. While SEMA3E was not detectable, we did observe SRSF7 expression widely across cell types, including neutrophils, monocytes, and B cells (Fig. 6f, g). SRSF7 plays a role in alternative RNA processing and mRNA export, but has no known role in aging or neurodegeneration. Further research will be required to determine if miRNAs like miR-6812-3p do indeed target SRSF7 in these specific cell types, and to uncover if this process contributes to the global decline of transcription observed with age.

Discussion

Our analysis of blood derived microRNAs provides insights into changes in microRNA abundance dependent on age, sex, and disease. While age clearly contributes to expression changes, sex has a more modest effect. In fact, most miRNAs show a similar behavior over the lifespan in males and females. This is generally in-line with recent results in transcriptomic mouse tissue aging[7,8]. Generally, our results compare well to other studies of miRNAs in aging[27], especially regarding miRNAs increasing with age, for which we observe high concordance. There are, however, miRNAs decreasing with age reported in the previous study for which we did not find evidence. The most extreme examples are miR-30d-5p and miR-505-5p, both increasing with age in our study in the healthy individuals. Nonetheless, given different cohorts with different ethnicity, varying age range, and distinct profiling technologies, we observed remarkable concordance between the studies.

Here, we observed that diseases globally disturb the normal aging progression of blood-borne miRNAs. While linear

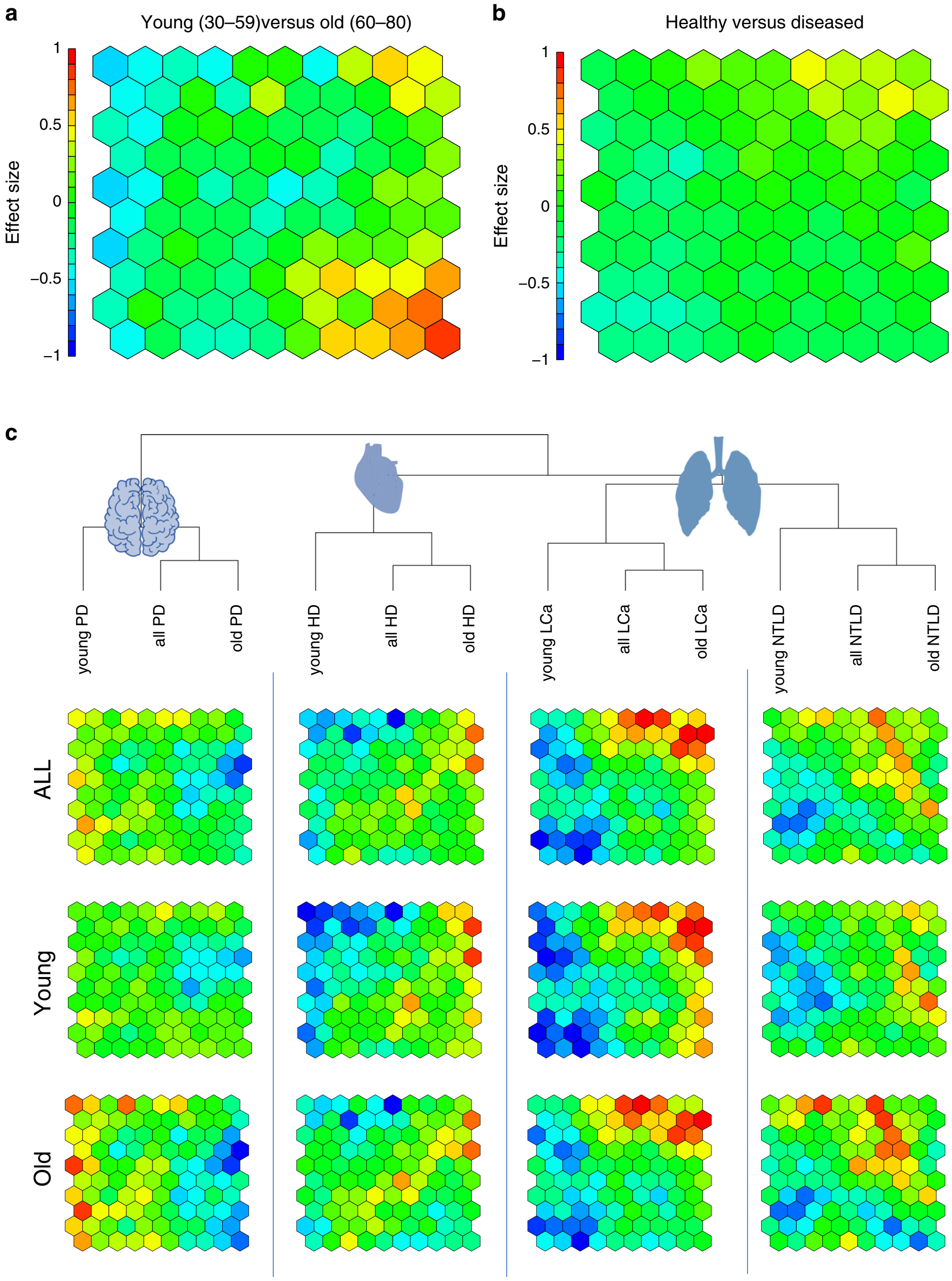


Fig. 4 Disease specificity of miRNA biomarkers. a Heat map representation of the SOM analysis as a 10 × 10 grid with 100 entries. Each cell contains at least one miRNA and up to 20 miRNAs. The full annotation of miRNAs to cells are provided in Supplementary Data 7). The cells are colored by the effect size of miRNAs for the comparison in old versus young. Red cells contain miRNAs with effect sizes >0.5 that are upregulated and in blue miRNAs that are downregulated with effect sizes <−0.5. **b** Same heat map as in **a** but colored for the difference in young versus old. The scale for the effect size has been kept the same as **a**. Thus fewer yellow/red, as well as blue spots indicate overall lower effect sizes. **c** Clustering of the SOM results in biomarkers for the four diseases and in all biomarkers independently of age, biomarkers for young patients and biomarker for old patients. The dendrogram has been computed from hierarchical clustering (complete linkage on the Euclidean distance). In all cases the biomarkers cluster by disease and not by age and the old biomarker set is closest to the all biomarker set while the young biomarker set has larger distances. Overall, NTLD and LCa markers are closest to each other, second closest are heart biomarkers and most different PD biomarkers. The SOM cells clearly highlight differences between biomarkers for diseases in young and old patients.

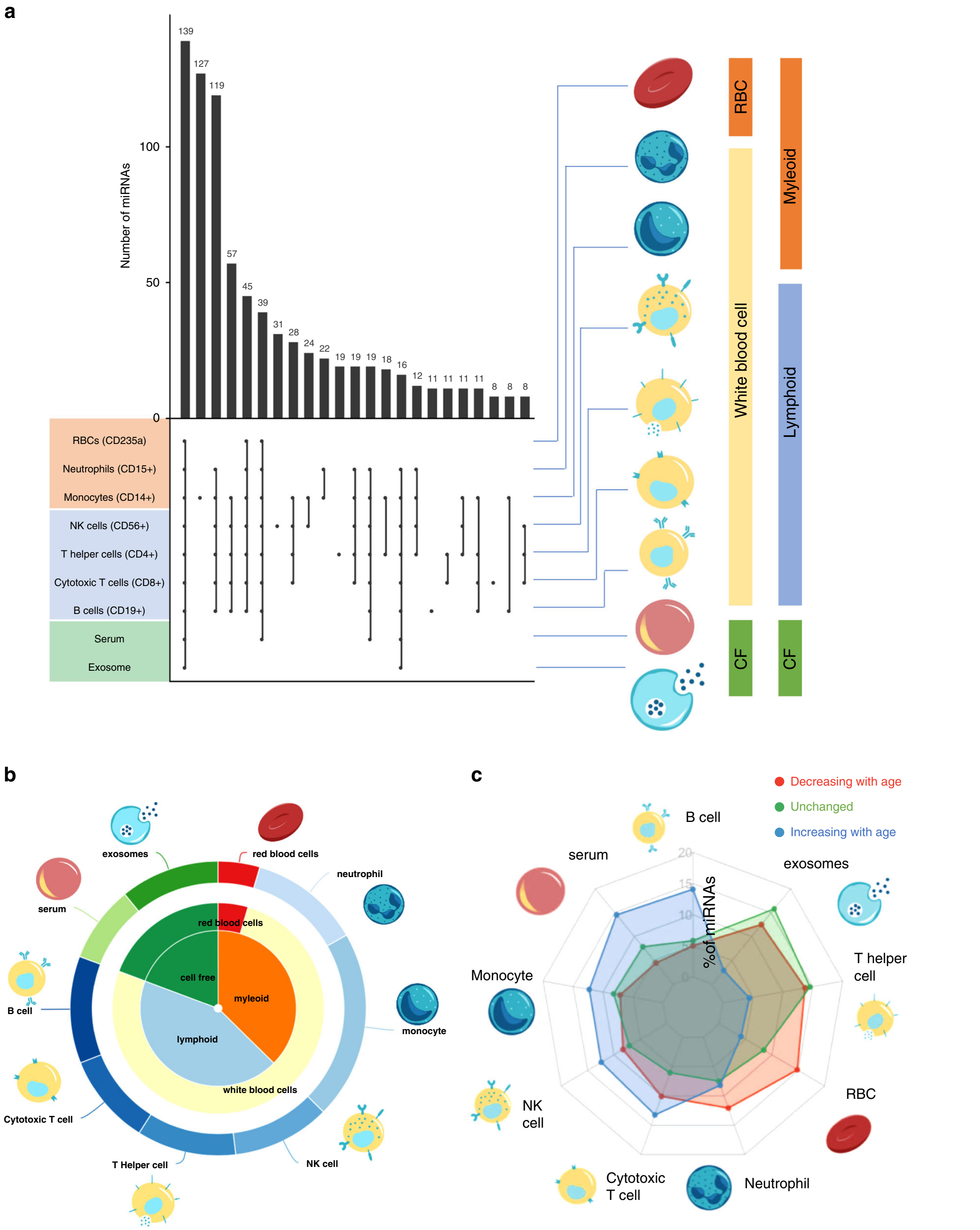
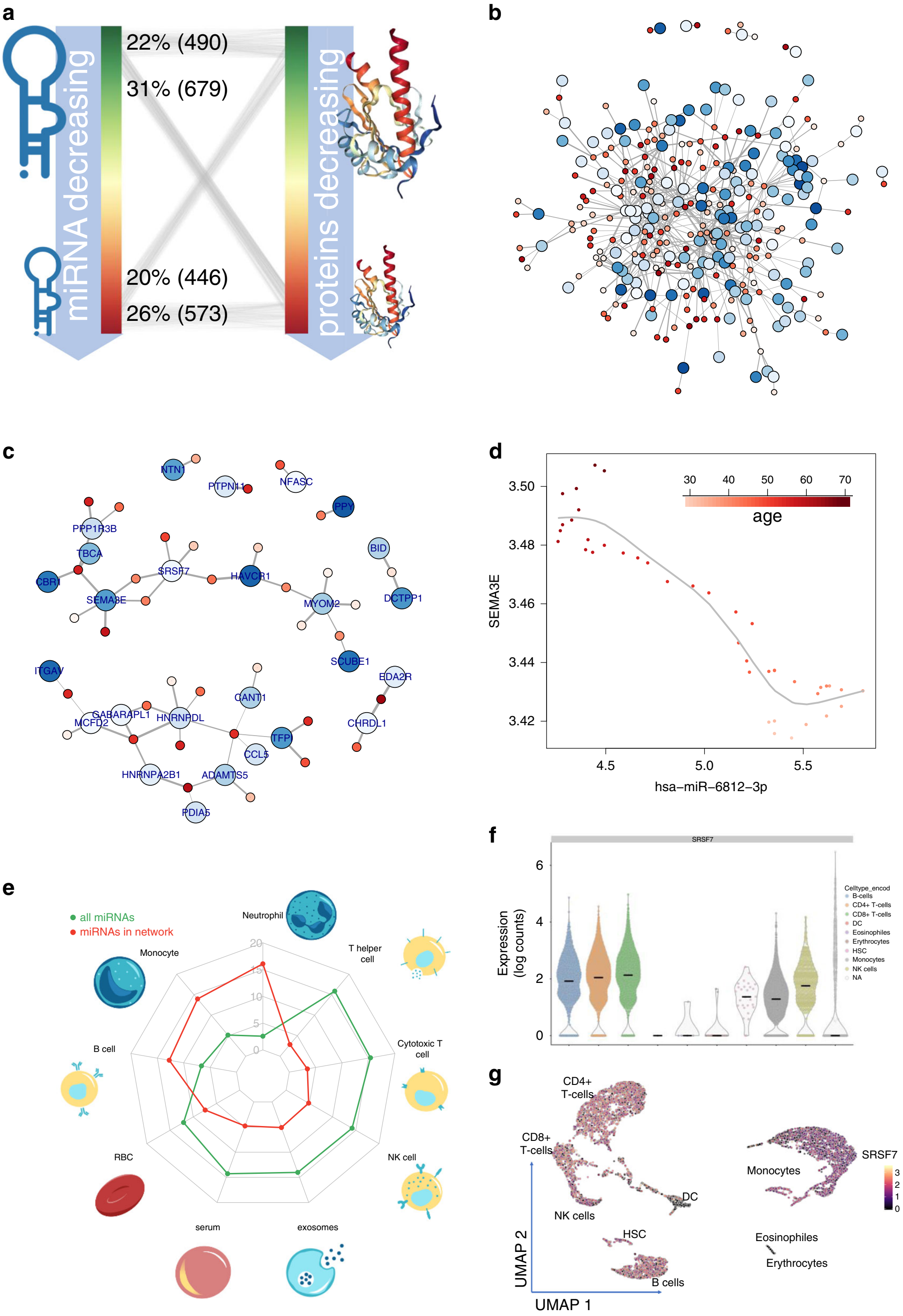


Fig. 5 Blood cell deconvolution. a The distribution of miRNAs in the different blood compounds. The rows are sorted by the blood compounds given on the right (RBC: red blood cell; CF: cell free), the columns are sorted according to a decreasing number of miRNAs. **b** Relative abundance of all miRNAs in the different blood compounds. **c** Distribution of miRNAs in cell types. The green distribution is the background and presents the relative composition of 1451 miRNAs in cluster 3. The blue distribution represents miRNAs increasing by age (cluster 4&5) and are enriched e.g., in B cells and serum. The red distribution represents miRNAs decreasing by age (cluster 1&2) and are enriched e.g., in neutrophils and RBCs.



modeling insufficiently explained changes with aging, distance correlation analysis identified 90 miRNAs that were decreasing and 26 that were increasing with age in a non-linear manner. These effects are, however, frequently not disease specific. If disease specific effects occur, they appear to establish themselves in given time windows throughout live. For example, lung and heart diseases show the largest effect sizes in the 4th to 5th decade of life, and Parkinson's disease showed the largest effect size in the 6th to 7th decade. All known biological factors including age, sex, and disease status together only explained part of the overall data variance. Thus, unknown biological variables and technical factors also contribute to miRNA abundance.

Our results underline not only the importance of age as a confounder in biomarker studies, but they show that age needs to be

Fig. 6 Age related miRNAs are correlated to age related proteins. a Correlation of miRNAs to proteins. miRNAs and proteins are sorted by increasing correlation with age. Thin lines are miRNA/gene interactions between top/bottom 10% of miRNAs and proteins. Numbers represent actual count of edges. **b**, **c** Core network. Proteins (larger nodes) are targeted by miRNAs (smaller nodes). Edge width correspond to the correlation. Blue nodes represent increase with age, red nodes decrease with age. The outer circles of the protein nodes indicate an expected an influence of the miRNAs leading to an increase with age. Panel **c** represents a more stringent version of the network from panel **b**. **d** One representative example of an edge from the network in **b**, **c**: SEMA3E and miR-6812-3p. Each dot represents all individuals in a time interval of 10 years, shifted between 30 and 70 years. SEMA3E is high expressed in older individuals while miR-6812-3p is low expressed (dark red points in the upper right corner). In young individuals the pattern is opposite (tale points in the lower right corner). **e** Blood cell compound distribution. miRNAs from the core network come from neutrophils, monocytes and B cells. **f** Violin plot of expression of SRSF7 in human blood cells. **g** UMAP embedding of human blood cells colored by expression of SRSF7.

incorporated into the definition of disease biomarkers. The age dependency of miRNA biomarkers may be even more prominent for acute diseases that are accompanied by drastic molecular changes. Furthermore, the influence of a disease on healthy aging miRNA patterns suggests that it is conceivable to define “negative biomarkers”, i.e., biomarkers that reflect the degree of disturbance of a given time-dependent pattern typically found in healthy individuals.

miRNAs comprise complex gene regulatory networks, and it is essential to identify the miRNA-targets that are regulated by a given miRNA network. However, this is already a demanding task for static networks, and it becomes even more challenging when considering how entire networks change with age. We attempted to overcome this complexity and identify a core miRNA network by implementing several stringent criteria: (i) the inclusion of miRNA-gene pairs only if experimental evidence exists, (ii) limiting the analysis to the top 5% of miRNAs decreasing with age, and (iii) the top 5% of proteins increasing with age and with pairwise absolute correlation of at least 0.6. This stringent parameter set identified a core network of 36 miRNAs and 26 proteins organized in two larger hubs with eight miRNAs targeting the axon guidance related semaphorin 3E (SEMA3E) and serine and arginine rich splicing factor 7 (SRSF7). Semaphorines play crucial roles during the development of the nervous system, especially in the hippocampal formation[35]. SEMA3E suppresses endothelial cell proliferation and angiogenic capacity, and in complex with PlexinD1 it inhibits recruitment of pericytes in endothelial cells[36]. Since we did not detect SEMA3E mRNA expression in single blood cell data we also explored other sources such as the Genotype-Tissue Expression (GTEx) project[37]. But also in the GTEx data no expression for the gene was reported in bulk sequencing data. It thus remains unclear how or if these miRNAs directly or indirectly impact SEMA3E protein levels in plasma. In this context, low abundant fractions of the blood such as exosomes might play a role. However, SRSF7, which belongs to a protein family linking alternative RNA processing to mRNA export[38], is expressed across a variety of circulating immune cells. This is intriguing as no role in aging or neurodegeneration is known.

Often, different technologies are available for high-throughput studies. To characterize the complete miRNome, usually microarrays or high-throughput sequencing are used. The choice of the best technology depends both, on technical factors and on the underlying biological question to be addressed. We decided to use microarray technology mostly because of the high dynamic range of blood miRNAs. In whole blood, the majority of reads (90–95%) are matching to few (2–5) miRNAs[39]. While generally a depletion is feasible[40], it bears the risk to alter the profile of other miRNAs especially since it has to be tailored for the respective sequencing technology. To use microarrays has however also disadvantages. MicroRNAs are often modified and build so-called isomiRs and basically all human miRNAs express different isoforms[41]. Likewise, data from the Rigoutsos lab demonstrate the importance and presence of isomiRs[42]. To address the age specific expression of isomiRs, single nucleotide resolution is required. Improved library preparation and sequencing methods together with increasing read numbers per sample will likely allow for an in-depth characterization of isomiRs in challenging specimens such as whole blood.

Another aspect for respective studies is the underlying specimen type. A literature search reveals that for human miRNA biomarker studies mostly plasma, serum, and blood cells (either PBMCs or whole blood) are considered with a more recent trend towards exosomes. Since we are interested in the connection of miRNA expression and the immune system by analyzing multiple diseases[43] we measured blood cells. Different aspects can be used to provide an even more comprehensive systemic picture of miRNAs and aging. First, the cell free part of the blood is also correlated to miRNA aging[44,45]. One important aspect are vesicles. Cellular senescence for example contributes to age-dependent changes in circulating extracellular vesicle cargo[46]. Moreover, the differential loading of vesicles is correlated to different human diseases[47–49]. Likewise, for the cellular part, resolution can be increased. For example, the miRNomes could be investigated per blood cell type[50]. One challenge is in that the purification of the different cell types by different isolation techniques potentially alters the miRNA content. Positive and negative selection, as well as Fluorescence-activated cell sorting (FACS) have a highly significant influence on the physiological miRNA content[32]. Here, single cell miRNA profiling might help to improve our understanding of age-related miRNA patterns in the future. At best, single cell miRNA data and cell free miRNA profiles are combined in the future using advancing sequencing technologies. Finally, such data might further our understanding of miRNAs in aging, diseases and their interplay with organ patterns that are only partially understood[29,51].

Over recent years, numerous studies have emerged highlighting systemic molecular aging factors detected with different omics technologies, including epigenetics, transcriptomics, and proteomics. Our study specifically extends our knowledge of blood and plasma-based miRNA patterns in aging. In our study we observe non-linear miRNA aging patterns. Moreover, the high degree of age-related biomarker patterns challenges the concept of age independent miRNA biomarker profiles, calling for different statistical models in aged and younger individuals. The changes with aging are not only attributed to one mature form, we also provide detailed insights into changes of the usage of the 3’ and 5’ mature arms in aging.

Furthering our understanding of age-related miRNA changes in healthy individuals and diseased patients will not only increase our understanding of age-related blood-borne gene regulation, but also improve miRNA-based biomarker development, and aid the development of RNA-based therapies.

Methods

Cohort. In this study, we processed data from $n_{\mathrm{total}} = 4433$ whole blood samples. We excluded 40 individuals (0.9%) because of insufficient data quality or missing clinical or demographic information. The final cohort consists thus of 4393 samples. These include unaffected controls ($n_{\mathrm{HC}} = 1{,}334$), Parkinson’s Disease ($n_{\mathrm{PD}} = 944$), heart diseases ($n_{\mathrm{HD}} = 607$), non-tumor lung diseases ($n_{\mathrm{NTLD}} = 586$), lung cancer ($n_{\mathrm{LC}} = 517$), and other diseases ($n_{\mathrm{OD}} = 405$). The diseases can be split further in sub-classes. For lung cancer, we included non-small cell, as well as small cell lung cancer. For non-small cell lung cancer, we can further divide them in adenocarcinoma and squamous cell carcinoma. These split in low grade and high-grade tumors according to the TNM grading. The lung cancer cohort has been previously described in more detail[52]. The heart diseases include coronary artery disease, dilated cardiomyopathies and acute coronary syndrome. The non-tumor lung diseases include mostly chronic obstructive pulmonary diseases, the other diseases include sepsis, liver cirrhosis, breast cancer, endometriosis, and melanoma patients. We aggregate the diseases to an organ level (heart, brain and lung). Only for the lung we split the cohort in cancer and non-cancer samples. This aggregation level has been selected in a manner to be able to distinguish between healthy and diseased aging by having sufficient cohort sizes. Detailed diagnoses for each sample are provided in Supplementary Data 1. All participants gave informed consent. The local ethics committee of Saarland University approved the study. The study has been conducted in compliance with all relevant ethical regulations regarding the use of human study participants.

RNA extraction and measurement of miRNAs. RNA from 4433 whole blood samples in PAXgeneTubes (BD Biosciences, Franklin Lakes, NJ, USA) was isolated using the PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany) using manufacturers recommendation. The extractions were done manually or semi-automatically on the Qiacube robot. The RNA was quantified using Nanodrop (Thermo Fisher Scientific, Waltham, MA, USA) and the RNA integrity was checked using a bioanalyzer with the RNA Nano Kit (Agilent Technologies, Santa Clara, CA, USA). The genome-wide expression profiles of human mature miRNAs was determined with Human miRNA microarrays and the miRNA Complete Labeling and Hyb Kit (Agilent Technologies). The labeled RNA was hybridized to the arrays for 20 h at 55 °C with 20 rpm rotation. The microarrays were subsequently washed twice, dried and scanned with 3 µm resolution in double-path mode (Agilent Technologies). The raw data were extracted using the manufacturers Feature Extraction software (Agilent Technologies). Details on the RNA extraction and microarray measurement procedure have been also previously described[53,54]. In difference to our previous studies we tried to further minimize any variability. In this study, we thus only included genome wide miRNA profiles that have been measured using the Agilent miRBase V21 biochip.

Blood cell deconvolution. To analyze the miRNA blood cell composition, we made use of our previous study that presented a high-resolution representation of human miRNAs in different blood compounds[50]. From the data, we asked which miRNAs are present in at least one sample of the respective blood compound and generated an upset plot from the data. In some detail, we included serum, microvesicles, red blood cells, CD15, CD19, CD8, CD56, CD4, and CD14 cells.

Correlation of age and sex to miRNAs. To find associations between the sex and the miRNA expression we applied 2-tailed non-parametric Wilcoxon Mann–Whitney tests. To compute linear correlation values between the age and miRNA expression values we computed the Pearson Correlation Coefficient (PC) and Spearman Correlation (SC). Further, to detect potentially non-linear relations between single miRNAs and the age we also computed the Distance Correlation (DI) between age and sex. To relate the DI and the SC, we computed a smoothed spline with eight degrees of freedom and computed the minimal Euclidean distance of each data point from the spline. Points with a distance of 0.02 (the threshold of 0.02 has been computed by a histogram-based approach) were highlighted and are considered to follow a non-linear trend with aging. In the further analyses, we applied only the rank-based Spearman Correlation (SC) instead of the Pearson Correlation that assumes linear effects in data. Beyond linear and non-linear correlations between single miRNAs and the age we also performed different standard dimension reduction technologies, including principal component analysis, t-stochastic neighborhood embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). To calculate the fraction of variance attributed to the age and sex we applied principal variant component analysis (PVCA), originally developed to discover batch effects in microarray experiments.

Analysis of arm shift events. Recently, we developed the miRSwitch database and analysis tool to identify and characterize human arm shift and arm switch events[30]. To detect associations between aging and differential arm usage we considered the following criteria. First, the percentage of the 5’ mature arm given the total expression of 3’ and 5’ arm must correlate with an absolute Spearman Correlation Coefficient > 0.2. Second, the correlation must reach a *p*-value of at least 0.05. The *p*-value is computed by the R cor.test function via the asymptotic t approximation. Third, the difference between the minimal and maximal percentage of 5’ arm expression for any samples must exceed 0.2 (20%). As fourth and last condition, the 3’ and 5’ mature form must have a different sign, i.e., the 5’ has to increase with age and the 3’ to decrease or vice versa. The miRNAs that were discovered by this procedure where then checked by miRSwitch.

Cluster analysis and miRNA enrichment analysis. We split the miRNAs in 5 groups, strongly decreasing with age (SC < −0.2), decreasing with age (SC between −0.2 and −0.1), not changing with age (SC between −0.1 and 0.1), miRNAs increasing with age (SC > 0.1 and <0.2) and miRNAs increasing strongly with age (SC > 0.2). For each cluster, we computed smoothed splines for each miRNA and the cluster average allowing three degrees of freedom. Further, we computed for disjoint age windows of five years whether miRNAs are significantly higher or lower in cases versus controls at an alpha level of 0.05 and colored them, respectively, in red and green. To find categories that are significantly enriched either for miRNAs increasing or decreasing over age we performed a miRNA enrichment analysis using the miEAA tool[55], which has been recently updated[56]. Thereby, for over 14,000 categories running sum statistics are computed. The sorted list of miRNAs (increasing correlation with age) is processed from left to right. Whenever a miRNA is located in a category the running sum is increased otherwise it is decreased. The running sum is then plotted along with 100 random permutation tests. Notably, the *p*-value is not computed from the permutations but exactly by using dynamic programming. A category showing a perfect “V” like shape would contain miRNAs that are increasing over age while a category following a pyramid like shape contains miRNAs that are decreasing over age.

Sliding window analysis based on Cohen’s d. Since *p*-values rely on the effect size and the cohort size different group sizes bias the results frequently. In our sliding window analysis, we observed substantial differences, i.e., cases and controls are not equally distributed across the age range. We thus performed all analyses using Cohen’s d as effect size. All effects with an absolute value of above 0.5 were considered relevant. Negative effect sizes thereby characterize downregulation and positive effect sizes upregulation. We computed effect sizes for each disease in windows of 10 years, shifted by one year, starting from 30 and ending at 70 years (i.e., the last window is from 70 to 79 years). Only when at least 20 cases and control measurements were available effect sizes were computed. The calculated effect sizes were then summarized and a smoothed spline with eight degrees of freedom were computed.

Self-organizing map (SOM) for finding disease patterns. One task in high dimensional data analysis is to group features and to generate lower dimensional representation of high dimensional data. Self-organizing maps (SOMs) are one type of artificial neural networks (ANNs), relying on competitive learning. As described by Kohonen already in 1982[57], in a network of adaptive elements “receiving signals from a primary event space, the signal representations are automatically mapped onto a set of output responses in such a way that the responses acquire the same topological order as that of the primary events”. From input data, a typically two-dimensional discretized representation of the input space is derived that can be visualized by heat maps. To compute self-organizing maps for patients and controls in an age dependent manner we computed the effect size for each disease group over all patients, for young patient (30–60 years) and for old patients (60–80 years) separately. Only 801 highest expressed miRNAs were included in this analysis. For the biomarker sets, a 10 × 10 hexagonal som grid was used to train a network. The data set was presented 10,000 times to the network. The learning rate was set to be between 0.05 and 0.01, meaning that the learning rate linearly decreased from 0.05 to 0.01 over the 10,000 iterations. To cluster the SOM cells, we performed hierarchical clustering. In more detail, we applied the R hclust function to carry out agglomerative complete linkage clustering. As distance measure we computed the Euclidean distance using the R dist function.

Plasma proteomics measurements. We used data from a recent study investigating the effect of aging on the human plasma proteome. In this study, 2925 proteins were measured using the SomaScan assay in 4264 subjects from the INTERVAL and LonGenity cohorts[5]. The SomaScan platform is based on modified single-stranded DNA aptamers binding to specific protein targets. Assay details were previously described. Relative Fluorescence Units (RFUs) were log10-transformed and we used a 10 years sliding window to estimate proteins trajectories throughout lifespan.

Target analysis and target network analysis. The main biological function of miRNAs is to bind the 3’ UTR of genes and to degrade the target mRNAs. In reality, miRNAs and genes thereby follow a n:m relation, i.e., one miRNA can regulate many genes and one gene is regulated by many miRNAs. Further, there exist different confidence levels to assume a pair-wise regulation of a miRNA to a target gene. Most relations are only predicted by one or several computational analyses. Another set is composed of miRNA gene pairs with weak evidence, e.g., from microarray experiments. The most reliable category consists of miRNA gene pairs with strong evidence, e.g., validated by reporter assays. We only considered this most reliable set of miRNA gene interactions and extracted the set from the miRTarBase database[34,58]. Our analysis highlighted that around 20% of miRNAs are increasing with age, 20% are decreasing and 60% are not age dependent. We assumed the same distribution for human plasma proteins changing with age and asked how many miRNAs going down with age regulate genes/proteins going up and down with age, respectively. Similarly, we asked how many miRNAs going up with age regulate genes/proteins going up and down with the age.

To construct a reliable core network, we combined five stringed filtering approaches and only considered those connections between miRNAs and genes that fulfill all filtering criteria. In the least stringent version the filters include (a) a strong experimental evidence of a target interaction from the literature; (b) one of the most decreasing miRNAs (5%) regulates (c) one of the most upregulated proteins (5%) over aging. To avoid a bias towards genes/proteins that are targeted only by one or few miRNAs, potentially also fragmenting the network, we (d) only considered proteins that are regulated by more than eight miRNAs. Next, we analyzed the correlation between miRNAs and genes/proteins in the network over 40 discrete age ranges from 30 to 70 years. Each age range thereby spans 10 years. For the 40 data points corresponding to 40 age windows we computed the Spearman correlation between miRNA expression in this age window and protein expression. As last criterion we added (e) only edges that have an absolute Spearman correlation of at least 0.6. This network has been visualized with the igraph library. Nodes were colored with respect to changes in age and edges weights relative to the absolute Spearman correlation.

Single cell analysis. We used data that have been made available by 10× genomics (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3). The profiles were subsequently processed with scater[59] and scran[60] with default parameters, cell type annotations with singleR[61].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw microarray measurements are freely available for any scientific purpose upon request as Excel Table and Tab Delimited Text file (110 MB) to data@ccb.uni-saarland.de. The use of the data for commercial purposes is prohibited.

Code availability

The data analysis has been performed using the R software for statistical computation (R 3.3.2 GUI 1.68 Mavericks build (7288)) using freely available packages. The following packages were used: ROC, RColorBrewer, preprocessCore, tsne, effsize, UpSetR, kohonen, fmsb, igraph. All packages are available from Bioconductor or CRAN.

Received: 18 February 2020; Accepted: 16 October 2020;
Published online: 24 November 2020

References

1. Harman, D. The aging process: major risk factor for disease and death. *Proc. Natl Acad. Sci. USA* **88**, 5360–5363 (1991).
2. Valdes, A. M., Glass, D. & Spector, T. D. Omics technologies and the study of human ageing. *Nat. Rev. Genet.* **14**, 601–607 (2013).
3. Deelen, J. et al. A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* **10**, 3669 (2019).
4. Aramillo Irizar, P. et al. Transcriptomic alterations during ageing reflect the shift from cancer to degenerative diseases in the elderly. *Nat. Commun.* **9**, 327 (2018).
5. Lehallier, B. et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
6. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
7. Schaum, N. et al. Ageing hallmarks exhibit organ-specific temporal signatures. *Nature* **583**, 596–602 (2020).
8. Tabula Muris, C. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
9. Hahn, O. et al. A nutritional memory effect counteracts benefits of dietary restriction in old mice. *Nat. Metab.* **1**, 1059–1073 (2019).
10. Villeda, S. A. et al. Young blood reverses age-related impairments in cognitive function and synaptic plasticity in mice. *Nat. Med.* **20**, 659–663 (2014).
11. Middeldorp, J. et al. Preclinical assessment of young blood plasma for Alzheimer disease. *JAMA Neurol.* **73**, 1325–1333 (2016).
12. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
13. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
14. Bushati, N. & Cohen, S. M. microRNA functions. *Annu. Rev. Cell Dev. Biol.* **23**, 175–205 (2007).
15. Gurtan, A. M. & Sharp, P. A. The role of miRNAs in regulating gene expression networks. *J. Mol. Biol.* **425**, 3582–3600 (2013).
16. Krek, A. et al. Combinatorial microRNA target predictions. *Nat. Genet.* **37**, 495–500 (2005).
17. Leidinger, P. et al. A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.* **14**, R78 (2013).
18. Keller, A. et al. Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement.* **12**, 565–576 (2016).
19. Keller, A. et al. Comprehensive analysis of microRNA profiles in multiple sclerosis including next-generation sequencing. *Mult. Scler.* **20**, 295–303 (2014).
20. Vogel, B. et al. Multivariate miRNA signatures as biomarkers for non-ischaemic systolic heart failure. *Eur. Heart J.* **34**, 2812–2822 (2013).
21. Smith-Vikos, T. & Slack, F. J. MicroRNAs and their roles in aging. *J. Cell Sci.* **125**, 7–17 (2012).
22. Somel, M. et al. MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res.* **20**, 1207–1218 (2010).
23. Drummond, M. J. et al. Aging and microRNA expression in human skeletal muscle: a microarray and bioinformatics analysis. *Physiol. Genomics* **43**, 595–603 (2011).
24. Zhang, H. et al. Investigation of microRNA expression in human serum during the aging process. *J. Gerontol. A Biol. Sci. Med. Sci.* **70**, 102–109 (2015).
25. Noren Hooten, N. et al. Age-related changes in microRNA levels in serum. *Aging* **5**, 725–740 (2013).
26. Meder, B. et al. Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin. Chem.* **60**, 1200–1208 (2014).
27. Huan, T. et al. Age-associated microRNA expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell* **17**, https://doi.org/10.1111/acel.12687 (2018).
28. Kehl, T. et al. miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz1022 (2019).
29. Ludwig, N. et al. Distribution of miRNA expression across human tissues. *Nucleic Acids Res.* **44**, 3865–3877 (2016).
30. Kern, F. et al. miRSwitch: detecting microRNA arm shift and switch events. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa323 (2020).
31. Chen, L. et al. miRNA arm switching identifies novel tumour biomarkers. *EBioMedicine* **38**, 37–46 (2018).
32. Schwarz, E. C. et al. Deep characterization of blood cell miRNomes by NGS. *Cell Mol. Life Sci.* **73**, 3169–3181 (2016).
33. Valiathan, R., Ashman, M. & Asthana, D. Effects of ageing on the immune system: infants to elderly. *Scand. J. Immunol.* **83**, 255–266 (2016).
34. Huang, H. Y. et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz896 (2019).
35. Gil, V. & Del Rio, J. A. Functions of plexins/neuropilins and their ligands during hippocampal development and neurodegeneration. *Cells* **8**, https://doi.org/10.3390/cells8030206 (2019).
36. Zhou, Y. F. et al. Sema3E/PlexinD1 inhibition is a therapeutic strategy for improving cerebral perfusion and restoring functional loss after stroke in aged rats. *Neurobiol. Aging* **70**, 102–116 (2018).
37. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
38. Muller-McNicoll, M. et al. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev.* **30**, 553–566 (2016).
39. Fehlmann, T. et al. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet.* **8**, 123 (2016).
40. Juzenas, S. et al. Depletion of erythropoietic miR-486-5p and miR-451a improves detectability of rare microRNAs in peripheral blood-derived small RNA sequencing libraries. *NAR Genom. Bioinform.* **2**, https://doi.org/10.1093/nargab/lqaa008 (2020).
41. Fehlmann, T. et al. A high-resolution map of the human small non-coding transcriptome. *Bioinformatics* **34**, 1621–1628 (2018).
42. Londin, E. et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl Acad. Sci. USA* **112**, E1106–E1115 (2015).
43. Keller, A. et al. Toward the blood-borne miRNome of human diseases. *Nat. Methods* **8**, 841–843 (2011).
44. Wang, H. et al. Transcriptome analysis of common and diverged circulating miRNAs between arterial and venous during aging. *Aging* **12**, 12987–13004 (2020).
45. Maffioletti, E. et al. miR-146a plasma levels are not altered in Alzheimer's disease but correlate with age and illness severity. *Front. Aging Neurosci.* **11**, 366 (2019).
46. Alibhai, F. J. et al. Cellular senescence contributes to age-dependent changes in circulating extracellular vesicle cargo and function. *Aging Cell* **19**, e13103 (2020).
47. Gomez, I. et al. Neutrophil microvesicles drive atherosclerosis by delivering miR-155 to atheroprone endothelium. *Nat. Commun.* **11**, 214 (2020).
48. Wei, Z. et al. Coding and noncoding landscape of extracellular RNA released by human glioma stem cells. *Nat. Commun.* **8**, 1145 (2017).
49. Cheng, M. et al. Circulating myocardial microRNAs from infarcted hearts are carried in exosomes and mobilise bone marrow progenitor cells. *Nat. Commun.* **10**, 959 (2019).
50. Juzenas, S. et al. A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Res.* **45**, 9290–9301 (2017).

51. Fehlmann, T., Ludwig, N., Backes, C., Meese, E. & Keller, A. Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol.* **13**, 1084–1088 (2016).
52. Fehlmann, T. et al. Evaluating the use of circulating MicroRNA profiles for lung cancer detection in symptomatic patients. *JAMA Oncol.* https://doi.org/10.1001/jamaoncol.2020.0001 (2020).
53. Keller, A. et al. Genome-wide MicroRNA expression profiles in COPD: early predictors for cancer development. *Genomics Proteom. Bioinform.* **16**, 162–171 (2018).
54. Ludwig, N. et al. Spring is in the air: seasonal profiles indicate vernal change of miRNA activity. *RNA Biol.* **16**, 1034–1043 (2019).
55. Backes, C., Khaleeq, Q. T., Meese, E. & Keller, A. miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.* **44**, W110–W116 (2016).
56. Kern, F. et al. miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa309 (2020).
57. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
58. Hsu, S. D. et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **39**, D163–D169 (2011).
59. McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
60. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).
61. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).

Acknowledgements

This work was supported by grants from the Michael J Fox Foundation (to Andreas Keller) and the Luxembourg National Research Fund (FNR) within the PEARL programme (FNR; FNR/P13/6682797 to Rejko Krüger).

Author contributions

T.F.: Data analysis, conception of the study and analyses; B.L.: Data analysis, manuscript drafting; N.S.: Data interpretation, manuscript drafting; O.H.: Data interpretation, manuscript drafting, data representation; M.K.: Data analysis; Y.L.: Data interpretation; N.G.: Data interpretation, data representation; L.G.: Data interpretation; C.B.: Data analysis; R.B.: Data interpretation, conception of the study and analyses; F.K.: Data analysis, data representation; R.K.: Data interpretation, conception of the study and analyses, providing clinical data and patient specimens; F.L.: Data interpretation, providing clinical data and patient specimens; N.L.: Performing analyses and contributing experimental data; B.M.: Data interpretation, conception of the study and analyses, providing clinical data and patient specimens; B.F.: Data interpretation, manuscript drafting; W.M.: Data interpretation; D.B.: Data interpretation; K.B.: Data interpretation; C.D.: Data interpretation; A.K.v.T.: Data interpretation, providing clinical data and patient specimens; G.W.E.: Data interpretation, providing clinical data and patient specimens; S.M.: Data interpretation, Performing analyses and contributing experimental data; N.B.: Data interpretation, Performing analyses and contributing experimental data; M.R.: Data interpretation, providing clinical data and patient specimens; T.W.C.: Data interpretation, manuscript drafting; E.M.: Data interpretation, conception of the study and analyses, manuscript drafting; A.K.: Data analysis, Data interpretation, conception of the study and analyses, manuscript drafting.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

M.K. is also employed by Hummingbird Diagnostic GmbH. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19665-1.

Correspondence and requests for materials should be addressed to A.K.

Peer review information *Nature Communications* thanks Lifang Hou and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4
Discussion and outlook

4.1 Discussion

In this thesis, we implemented tools and resources for miRNA research, investigated technological advances and different sources of bias, and showed the potential of miRNAs as non-invasive biomarkers. The developed tools and resources belong to the current state-of-the art in the miRNA field and are regularly used by researchers over the entire world, as shown by their usage statistics in Figure 4.1. Nevertheless, limitations to the herein presented work exist and further improvements can be made.

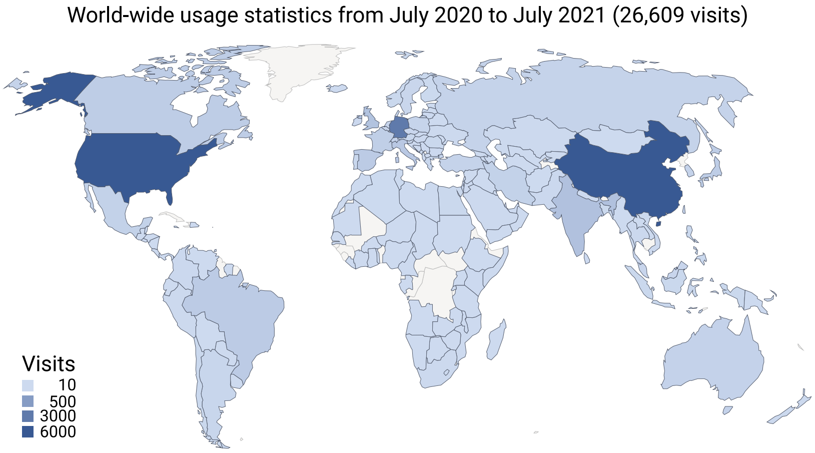

World-wide usage statistics from July 2020 to July 2021 (26,609 visits)
Visits
10
500
3000
6000

Figure 4.1: World-wide usage statistics of the tools and resources presented in this thesis, showing the number of visits from each country since July 2020. These results exclude the visits of miRPathDB 1 and 2 since no data was collected for it.

To begin with, while many herein described studies characterized other non-coding RNAs in addition to miRNAs, the latter were always the major focus of our studies. One reason was that developing the same level of skills for all other non-coding RNA classes is by far beyond the scope of a single thesis and even of one workgroup. Nevertheless, more detailed analyses of these other RNA classes could lead to additional insights, especially since they start to gain more attention also with respect to their involvement in diseases [120, 259, 507]. In particular, the regulatory role of circular RNAs acting as miRNA sponges could lead to complementary knowledge and possibly contribute to biomarker signatures [237, 508].

One aspect that was evaluated in multiple publications of this thesis was the identification of novel miRNA candidates from NGS data. While substantial improvements could be reached with respect to the prediction accuracy and algorithm scalability [1], this process is still strongly affected by false positives and requires extensive experimental validation [4, 6, 7]. Yet, high-throughput validation methods are still not reliable enough, calling for a targeted validation of small candidate sets [7, 12]. It is mandatory to keep in mind that even low-throughput validation methods have their limitations, e.g., artificial *in vitro* overexpression might lead to incorrect assumptions about *in vivo* physiological situations. Another issue faced with miRNA candidate prediction is the continuous re-discovery of miRNA candidates. We addressed this aspect with the creation of a comprehensive miRNA candidate database, miRCarta [5]. However our focus lied on human data, and thus this aspect still persists for other species. Nevertheless, even for human data, improvements could be made, by integrating the continuously growing amount of sequencing data, thus leading to a better coverage of potential miRNA candidates. Similar observations as for the miRNA prediction field hold true for the miRNA target prediction tools. Predicted targets must be evaluated with care because of their high false positive rates and only thorough experimental validation can shed light on real interactions [288]. In particular, all high confidence methods are low-throughput methods, although efforts are pursued to augment their scalability by combining them with automatic processing [169]. This results in a validation imbalance observed in the targeting field, since mainly well-known miRNAs and targets expected to be involved in diseases, especially in cancer, are being investigated [459]. In addition, experimental validation in species other than human or mouse is inherently under-represented [290]. One core issue with this problematic is that it leads to a bias in all analyses building up on experimentally validated targets, including functional enrichment analyses. Although we implemented computational approaches to reduce the impact of the existing validation imbalance, it cannot be fully excluded. Therefore, more reliable high-throughput methods that allow unbiased evaluations are necessary.

Next, the evaluation of different methods and approaches to measure miRNAs revealed that while miRNA profiles were often highly reproducible, the bias introduced by different technologies or library preparation methods is substantial [13, 15–17]. This was also highlighted by another multi-center study, which found that the often-observed ligation bias could be reduced by the addition of degenerate bases to the sequencing adapters [373]. In addition, not all libraries are able to capture all occurring RNA modifications, such as 5′ capping and 3′ phosphorylation, and thereby shift the landscape of considered RNAs [16, 509]. This aspect needs to be considered when integrating datasets of different sources, as well as when quantifying the diversity of certain RNAs. It also directly influences the prediction of miRNA candidates from NGS data [13].

Furthermore, while the herein evaluated biomarker studies show promising results, their translation to the clinic still needs substantial additional efforts. One aspect is the discussed technological bias, hampering the validation of detected miRNA signatures. Indeed, many studies often fail to validate their findings in independent cohorts [337, 339]. In addition, miRNA signature profiles can be affected by other factors such as age [26], sex [342], ethnic background [254], and even seasonal changes [10]. Moreover, since whole-blood is composed of a mix of different cell populations as well as cell-free components, spanning erythrocytes, leucocytes, thrombocytes, and plasma, whole-blood miRNA profiles can be affected by blood cell composition changes [510, 511]. Furthermore, generalization might fail because of a population bias present in the evaluated cohorts, possibly linked with different diagnostic practices implemented in different clinics. The effect of drug treatments on miRNA profiles is still only poorly characterized and might represent an additional confounding factor [512–514]. Moreover, the evaluation of retrospective studies might overestimate the actual performance of a biomarker signature in the clinic, and thus large prospective studies need to be conducted to validate them [25]. In addition, the biomarker signature needs to be evaluated together with current clinical methods to measure the provided benefit. In particular when considering prospective studies, the use of miRNAs that are generally deregulated in various diseases, such as hsa-miR-144-5p and hsa-miR-21-5p is likely to negatively impact the performance [340, 341]. The resulting shift in cohort composition might also lead to issues with machine learning algorithms trained on a retrospective cohort with different composition. Algorithms that scale well to differently composed cohorts are still actively researched and need further tuning[515, 516].

4.2 *Outlook*

While most RNA sequencing studies have been performed on pools of large populations of input cells, one major weakness of this approach is that no distinction between the cell types within the same sample is possible. In the last decade, breakthrough technologies have enabled the characterization of mRNA at an unprecedented scale, enabling the profiling of thousands of single cells per sample [517–520]. Consequently, single-cell atlases have been created, allowing a detailed analysis of cellular states and regulatory factors of neurodegenerative diseases in brain regions [521–524], as well as an investigation of the development of cancer cell subtypes of lung cancer patients [525]. Although single-cell miRNA sequencing methods are being developed, they are still restricted in their throughput and resolution, since each cell must be processed individually and contains only low amounts of miRNA [492]. Thus, the development of a method allowing to measure miRNAs at a similar scale than mRNAs would be a major breakthrough leading to new insights into the distribution and regulatory role of miRNAs in different cell types. Alternatively, reliable predictions of miRNA levels from single-cell mRNA data could present an interesting possibility to take advantage of the breakthroughs in this field. Although respective approaches are emerging, they are not yet suitable for a general application [526].

While multi-omics datasets profiling miRNA expression levels in addition to other omic types become increasingly popular, they mostly focus on mixes of cell populations [527–529]. Recently, methods have emerged that demonstrate the feasibility of multi-omics measurements at the single-cell level characterizing for example mRNAs together with chromatin accessibility profiles [530], mRNAs together with selected proteins [531], as well as all three omic types simultaneously [532]. Similar trends are pursued for miRNA-sequencing data, where methods capturing mRNAs and miRNAs of the same cell have been evaluated [533, 534]. However, these methods are currently limited in their scalability and sensitivity, with only a fraction of the expected number of miRNAs being detected. Nevertheless, more mature datasets promise additional insights into the regulatory mechanisms of miRNAs, and will likely enable the creation of more accurate miRNA target prediction tools.

Another technique that has recently gained attention is thiol(SH)-linked alkylation for the metabolic sequencing (SLAM-seq). It was first developed in the context of RNA sequencing [535] and recently adapted for small RNAs. This technique allows to measure the turnover of miRNAs in living cells, and was used to study the kinetics of miRNA biogenesis in *Drosophila* [536]. It opens new possibilities to study the causes of deregulated miRNA expression patterns in diseases.

With large public sequencing initiatives such as the 100,000 Genomes Project [537], the number of characterized human genomes and identified variants is quickly growing. Therefore, new data structures and algorithms that account for known individual variations were developed [538–540]. While these developments are not yet integrated into miRNA target prediction tools or tools estimating the influence of SNVs on the targetome, it is likely that they represent an important step to reach scalable solutions.

In conclusion, the rapid development of new experimental methods and the generation of increasingly complex datasets calls for the implementation of a novel generation of efficient software tools allowing to analyze and integrate all new aspects in the coming years.

Bibliography

[1] Fehlmann T, Backes C, Kahraman M, Haas J, Ludwig N, Posch AE, Würstle ML, Hübenthal M, Franke A, Meder B, Meese E, Keller A (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res*, 45:8731–8744

[2] Fehlmann T, Kern F, Laham O, Backes C, Solomon J, Hirsch P, Volz C, Müller R, Keller A (2021) miRMaster 2.0: multi-species non-coding RNA sequencing analyses at scale. *Nucleic Acids Res*, 49:W397–W408, W1

[3] Kern F, Amand J, Senatorov I, Isakova A, Backes C, Meese E, Keller A, Fehlmann T (2020) miRSwitch: detecting microRNA arm shift and switch events. *Nucleic Acids Res*, 48:W268–W274, W1

[4] Fehlmann T, Backes C, Alles J, Fischer U, Hart M, Kern F, Langseth H, Rounge T, Umu SU, Kahraman M, Laufer T, Haas J, Staehler C, Ludwig N, Hübenthal M, Meder B, Franke A, Lenhof HP, Meese E, Keller A (2018) A high-resolution map of the human small non-coding transcriptome. *Bioinformatics*, 34:1621–1628

[5] Backes C, Fehlmann T, Kern F, Kehl T, Lenhof HP, Meese E, Keller A (2018) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res*, 46:D160–D167, D1

[6] Fehlmann T, Laufer T, Backes C, Kahramann M, Alles J, Fischer U, Minet M, Ludwig N, Kern F, Kehl T, Galata V, Düsterloh A, Schrörs H, Kohlhaas J, Bals R, Huwer H, Geffers L, Krüger R, Balling R, Lenhof HP, Meese E, Keller A (2019) Large-scale validation of miRNAs by disease association, evolutionary conservation and pathway activity. *RNA Biol*, 16:93–103

[7] Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, Hart M, Abu-Halima M, Grässer FA, Lenhof HP, Keller A, Meese E (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res*, 47:3353–3364

[8] Ludwig N, Leidinger P, Becker K, Backes C, Fehlmann T, Pallasch C, Rheinheimer S, Meder B, Stähler C, Meese E, Keller A (2016) Distribution of miRNA expression across human tissues. *Nucleic Acids Res*, 44:3865–3877

[9] Fehlmann T, Ludwig N, Backes C, Meese E, Keller A (2016) Distribution of microRNA biomarker candidates in solid tissues and body fluids. *RNA Biol*, 13:1084–1088

[10] Ludwig N, Hecksteden A, Kahraman M, Fehlmann T, Laufer T, Kern F, Meyer T, Meese E, Keller A, Backes C (2019) Spring is in the air: seasonal profiles indicate vernal change of miRNA activity. *RNA Biol*, 16:1034–1043

[11] Fehlmann T, Backes C, Pirritano M, Laufer T, Galata V, Kern F, Kahraman M, Gasparoni G, Ludwig N, Lenhof HP, Gregersen HA, Francke R, Meese E, Simon M, Keller A (2019) The sncRNA Zoo: a repository for circulating small noncoding RNAs in animals. *Nucleic Acids Res*, 47:4431–4441

[12] Isakova A, Fehlmann T, Keller A, Quake SR (2020) A mouse tissue atlas of small noncoding RNA. *Proc Natl Acad Sci U S A*, 117:25634–25645

[13] Fehlmann T, Reinheimer S, Geng C, Su X, Drmanac S, Alexeev A, Zhang C, Backes C, Ludwig N, Hart M, An D, Zhu Z, Xu C, Chen A, Ni M, Liu J, Li Y, Poulter M, Li Y, Stähler C, Drmanac R, Xu X, Meese E, Keller A (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics*, 8:123

[14] Ludwig N, Fehlmann T, Galata V, Franke A, Backes C, Meese E, Keller A (2018) Small ncRNA-Seq Results of Human Tissues: Variations Depending on Sample Integrity. *Clin Chem*, 64:1074–1084

[15] Pirritano M, Fehlmann T, Laufer T, Ludwig N, Gasparoni G, Li Y, Meese E, Keller A, Simon M (2018) Next Generation Sequencing Analysis of Total Small Noncoding RNAs from Low Input RNA from Dried Blood Sampling. *Anal Chem*, 90:11791–11796

[16] Meistertzheim M, Fehlmann T, Drews F, Pirritano M, Gasparoni G, Keller A, Simon M (2019) Comparative Analysis of Biochemical Biases by Ligation- and Template-Switch-Based Small RNA Library Preparation Protocols. *Clin Chem*, 65:1581–1591

[17] Li Y, Fehlmann T, Borcherding A, Drmanac S, Liu S, Groeger L, Xu C, Callow M, Villarosa C, Jorjorian A, Kern F, Grammes N, Meese E, Jiang H, Drmanac R, Ludwig N, Keller A (2021) CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing. *Nucleic Acids Res*, 49:e10

[18] Hamberg M, Backes C, Fehlmann T, Hart M, Meder B, Meese E, Keller A (2016) MiRTargetLink–miRNAs, Genes and Interaction Networks. *Int J Mol Sci*, 17:564

[19] Backes C, Kehl T, Stöckel D, Fehlmann T, Schneider L, Meese E, Lenhof HP, Keller A (2017) miRPathDB: a new dictionary on microRNAs and target pathways. *Nucleic Acids Res*, 45:D90–D96, D1

[20] Kehl T, Kern F, Backes C, Fehlmann T, Stöckel D, Meese E, Lenhof HP, Keller A (2020) miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic Acids Res*, 48:D142–D147, D1

[21] Fehlmann T, Sahay S, Keller A, Backes C (2019) A review of databases predicting the effects of SNPs in miRNA genes or miRNA-binding sites. *Brief Bioinform*, 20:1011–1020

[22] Kern F, Fehlmann T, Solomon J, Schwed L, Grammes N, Backes C, Van Keuren-Jensen K, Craig DW, Meese E, Keller A (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res*, 48:W521–W528, W1

[23] Ludwig N, Fehlmann T, Kern F, Gogol M, Maetzler W, Deutscher S, Gurlit S, Schulte C, von Thaler AK, Deuschle C, Metzger F, Berg D, Suenkel U, Keller V, Backes C, Lenhof HP, Meese E, Keller A (2019) Machine Learning to Detect Alzheimer's Disease from Circulating Non-coding RNAs. *Genomics Proteomics Bioinformatics*, 17:430–440

[24] Kern F, Fehlmann T, Violich I, Alsop E, Hutchins E, Kahraman M, Grammes NL, Guimarães P, Backes C, Poston KL, Casey B, Balling R, Geffers L, Krüger R, Galasko D, Mollenhauer B, Meese E, Wyss-Coray T, Craig DW, Van Keuren-Jensen K, Keller A (2021) Deep sequencing of sncRNAs reveals hallmarks and regulatory modules of the transcriptome during Parkinson's disease progression. *Nat Aging*, 1:309–322

[25] Fehlmann T, Kahraman M, Ludwig N, Backes C, Galata V, Keller V, Geffers L, Mercaldo N, Hornung D, Weis T, Kayvanpour E, Abu-Halima M, Deuschle C, Schulte C, Suenkel U, von Thaler AK, Maetzler W, Herr C, Fähndrich S, Vogelmeier C, Guimaraes P, Hecksteden A, Meyer T, Metzger F, Diener C, Deutscher S, Abdul-Khaliq H, Stehle I, Haeusler S, Meiser A, Groesdonk HV, Volk T, Lenhof HP, Katus H, Balling R, Meder B, Kruger R, Huwer H, Bals R, Meese E, Keller A (2020) Evaluating the Use of Circulating MicroRNA Profiles for Lung Cancer Detection in Symptomatic Patients. *JAMA Oncol*, 6:714–723

[26] Fehlmann T, Lehallier B, Schaum N, Hahn O, Kahraman M, Li Y, Grammes N, Geffers L, Backes C, Balling R, Kern F, Krüger R, Lammert F, Ludwig N, Meder B, Fromm B, Maetzler W, Berg D, Brockmann K, Deuschle C, von Thaler AK, Eschweiler GW, Milman S, Barziliai N, Reichert M, Wyss-Coray T, Meese E, Keller A (2020) Common diseases alter the physiological age-related blood microRNA profile. *Nat Commun*, 11:5958

[27] Roser M, Ortiz-Ospina E, Ritchie H. Life Expectancy. 2013. URL: `https://ourworldindata.org/life-expectancy` (visited on 06/20/2021)

[28] Jones DS, Podolsky SH, Greene JA (2012) The Burden of Disease and the Changing Task of Medicine. *N Engl J Med*, 366:2333–2338

[29] WHO (2020) Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. *World Health Organization*

[30] Patterson C, International AD (2018) World Alzheimer report 2018. Alzheimer's Disease International

[31] Yang W, Hamilton JL, Kopil C, Beck JC, Tanner CM, Albin RL, Ray Dorsey E, Dahodwala N, Cintina I, Hogan P, Thompson T (2020) Current and projected future economic burden of Parkinson's disease in the U.S. *NPJ Parkinsons Dis*, 6:15

[32] Horn L, Lovly CM (2018) Neoplasms of the Lung. Jameson JL, Fauci AS, Kasper DL, Hauser SL, Longo DL, Loscalzo J, editors, *Harrison's Principles of Internal Medicine*. McGraw-Hill Education, New York, NY

[33] Brierly JD, Gospodarowicz MK, Wittekind C, editors (2016) TNM Classification of Malignant Tumours. 8th edition. Wiley-Blackwell

[34] Navada S, Lai P, Schwartz AG, Kalemkerian GP (2006) Temporal trends in small cell lung cancer: Analysis of the national Surveillance, Epidemiology, and End-Results (SEER) database. *JCO*, 24:7082–7082

[35] Chalela R, Curull V, Enríquez C, Pijuan L, Bellosillo B, Gea J (2017) Lung adenocarcinoma: from molecular basis to genome-guided therapy and immunotherapy. *J Thorac Dis*, 9:2142–2158

[36] Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, Sougnez C, Auclair D, Lawrence M, Stojanov P, Cibulskis K, Choi K, de Waal L, Sharifnia T, Brooks A, Greulich H, Banerji S, Zander T, Seidel D, Leenders F, Ansén S, Ludwig C, Engel-Riedel W, Stoelben E, Wolf J, Goparju C, Thompson K, Winckler W, Kwiatkowski D, Johnson BE, Jänne PA, Miller VA, Pao W, Travis WD, Pass H, Gabriel S, Lander E, Thomas RK, Garraway LA, Getz G, Meyerson M (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150:1107–1120

[37] Office of the Surgeon General (US), Office on Smoking and Health (US) (2004) The Health Consequences of Smoking: A Report of the Surgeon General. Centers for Disease Control and Prevention (US), Atlanta (GA)

[38] Cancer of the Lung and Bronchus - Cancer Stat Facts. SEER. URL: `https://seer.cancer.gov/statfacts/html/lungb.html` (visited on 06/19/2021)

[39] National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*, 365:395–409

[40] Brussino L, Culla B, Bucca C, Giobbe R, Boita M, Isaia G, Heffler E, Oliaro A, Filosso P, Rolla G (2014) Inflammatory cytokines and VEGF measured in exhaled breath condensate are correlated with tumor mass in non-small cell lung cancer. *J Breath Res*, 8:027110

[41] Mazzone PJ, Wang XF, Lim S, Jett J, Choi H, Zhang Q, Beukemann M, Seeley M, Martino R, Rhodes P (2015) Progress in the development of volatile exhaled breath signatures of lung cancer. *Ann Am Thorac Soc*, 12:752–757

[42] Kim YJ, Sertamo K, Pierrard MA, Mesmin C, Kim SY, Schlesser M, Berchem G, Domon B (2015) Verification of the biomarker candidates for non-small-cell lung cancer using a targeted proteomics approach. *J Proteome Res*, 14:1412–1419

[43] Ma S, Wang W, Xia B, Zhang S, Yuan H, Jiang H, Meng W, Zheng X, Wang X (2016) Multiplexed Serum Biomarkers for the Detection of Lung Cancer. *EBioMedicine*, 11:210–218

[44] Montani F, Marzi MJ, Dezi F, Dama E, Carletti RM, Bonizzi G, Bertolotti R, Bellomi M, Rampinelli C, Maisonneuve P, Spaggiari L, Veronesi G, Nicassio F, Di Fiore PP, Bianchi F (2015) miR-Test: a blood test for lung cancer early detection. *J Natl Cancer Inst*, 107:djv063

[45] Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A, Hruban RH, Wolfgang CL, Goggins MG, Dal Molin M, Wang TL, Roden R, Klein AP, Ptak J, Dobbyn L, Schaefer J, Silliman N, Popoli M, Vogelstein JT, Browne JD, Schoen RE, Brand RE, Tie J, Gibbs P, Wong HL, Mansfield AS, Jen J, Hanash SM, Falconi M, Allen PJ, Zhou S, Bettegowda C, Diaz LA, Tomasetti C, Kinzler KW, Vogelstein B, Lennon AM, Papadopoulos N (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359:926–930

[46] Abbosh C et al. (2017) Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, 545:446–451

[47] Cao C, Manganas C, Ang SC, Peeceeyen S, Yan TD (2013) Video-assisted thoracic surgery versus open thoracotomy for non-small cell lung cancer: a meta-analysis of propensity score-matched patients. *Interact Cardiovasc Thorac Surg*, 16:244–249

[48] Cao C, Frick AE, Ilonen I, McElnay P, Guerrera F, Tian DH, Lim E, Rocco G (2018) European questionnaire on the clinical use of video-assisted thoracoscopic surgery. *Interact Cardiovasc Thorac Surg*, 27:379–383

[49] Videtic GMM, Donington J, Giuliani M, Heinzerling J, Karas TZ, Kelsey CR, Lally BE, Latzka K, Lo SS, Moghanaki D, Movsas B, Rimner A, Roach M, Rodrigues G, Shirvani SM, Simone CB, Timmerman R, Daly ME (2017) Stereotactic body radiation therapy for early-stage non-small cell lung cancer: Executive Summary of an ASTRO Evidence-Based Guideline. *Pract Radiat Oncol*, 7:295–301

[50] Guckenberger M, Andratschke N, Dieckmann K, Hoogeman MS, Hoyer M, Hurkmans C, Tanadini-Lang S, Lartigau E, Romero AM, Senan S, Verellen D (2017) ESTRO ACROP consensus guideline on implementation and practice of stereotactic body radiotherapy for peripherally located early stage non-small cell lung cancer. *Radiother Oncol*, 124:11–17

[51] Paz-Ares L, Mezger J, Ciuleanu TE, Fischer JR, von Pawel J, Provencio M, Kazarnowicz A, Losonczy G, de Castro G, Szczesna A, Crino L, Reck M, Ramlau R, Ulsperger E, Schumann C, Miziara JEA, Lessa ÁE, Dediu M, Bálint B, Depenbrock H, Soldatenkova V, Kurek R, Hirsch FR, Thatcher N, Socinski MA, INSPIRE investigators (2015) Necitumumab plus pemetrexed and cisplatin as first-line therapy in patients with stage IV non-squamous non-small-cell lung cancer (INSPIRE): an open-label, randomised, controlled phase 3 study. *Lancet Oncol*, 16:328–337

[52] Lindeman NI, Cagle PT, Aisner DL, Arcila ME, Beasley MB, Bernicker EH, Colasacco C, Dacic S, Hirsch FR, Kerr K, Kwiatkowski DJ, Ladanyi M, Nowak JA, Sholl L, Temple-Smolkin R, Solomon B, Souter LH, Thunnissen E, Tsao MS, Ventura CB, Wynes MW, Yatabe Y (2018) Updated Molecular Testing Guideline for the Selection of Lung Cancer Patients for Treatment With Targeted Tyrosine Kinase Inhibitors: Guideline From the College of American Pathologists, the International Association for the Study of Lung Cancer, and the Association for Molecular Pathology. *J Thorac Oncol*, 13:323–358

[53] Nakagawa K et al. (2019) Ramucirumab plus erlotinib in patients with untreated, EGFR-mutated, advanced non-small-cell lung cancer (RELAY): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol*, 20:1655–1669

[54] Herbst RS, Giaccone G, de Marinis F, Reinmuth N, Vergnenegre A, Barrios CH, Morise M, Felip E, Andric Z, Geater S, Özgüroğlu M, Zou W, Sandler A, Enquist I, Komatsubara K, Deng Y, Kuriki H, Wen X, McCleland M, Mocci S, Jassem J, Spigel DR (2020) Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC. *N Engl J Med*, 383:1328–1339

[55] Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, Gottfried M, Peled N, Tafreshi A, Cuffe S, O'Brien M,

Rao S, Hotta K, Leiby MA, Lubiniecki GM, Shentu Y, Rangwala R, Brahmer JR (2016) Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *N Engl J Med*, 375:1823–1833

[56] Hellmann MD, Ciuleanu TE, Pluzanski A, Lee JS, Otterson GA, Audigier-Valette C, Minenza E, Linardou H, Burgers S, Salman P, Borghaei H, Ramalingam SS, Brahmer J, Reck M, O'Byrne KJ, Geese WJ, Green G, Chang H, Szustakowski J, Bhagavatheeswaran P, Healey D, Fu Y, Nathan F, Paz-Ares L (2018) Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *N Engl J Med*, 378:2093–2104

[57] DeTure MA, Dickson DW (2019) The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegener*, 14:32

[58] Hashimoto M, Rockenstein E, Crews L, Masliah E (2003) Role of protein aggregation in mitochondrial dysfunction and neurodegeneration in Alzheimer's and Parkinson's diseases. *Neuromolecular Med*, 4:21–36

[59] Goedert M, Klug A, Crowther RA (2006) Tau protein, the paired helical filament and Alzheimer's disease. *J Alzheimers Dis*, 9:195–207

[60] Lin MT, Beal MF (2006) Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature*, 443:787–795

[61] Nixon RA (2007) Autophagy, amyloidogenesis and Alzheimer disease. *J Cell Sci*, 120:4081–4091

[62] Lindholm D, Wootz H, Korhonen L (2006) ER stress and neurodegenerative diseases. *Cell Death Differ*, 13:385–392

[63] Heppner FL, Ransohoff RM, Becher B (2015) Immune attack: the role of inflammation in Alzheimer disease. *Nat Rev Neurosci*, 16:358–372

[64] Petersen RC (2018) How early can we diagnose Alzheimer disease (and is it sufficient)?: The 2017 Wartenberg lecture. *Neurology*, 91:395–402

[65] Jack CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, Holtzman DM, Jagust W, Jessen F, Karlawish J, Liu E, Molinuevo JL, Montine T, Phelps C, Rankin KP, Rowe CC, Scheltens P, Siemers E, Snyder HM, Sperling R, Elliott C, Masliah E, Ryan L, Silverberg N (2018) NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement*, 14:535–562

[66] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH (2011) Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on

Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7:280–292

[67] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7:270–279

[68] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, 7:263–269

[69] Mattsson N, Zetterberg H, Hansson O, Andreasen N, Parnetti L, Jonsson M, Herukka SK, van der Flier WM, Blankenstein MA, Ewers M, Rich K, Kaiser E, Verbeek M, Tsolaki M, Mulugeta E, Rosén E, Aarsland D, Visser PJ, Schröder J, Marcusson J, de Leon M, Hampel H, Scheltens P, Pirttilä T, Wallin A, Jönhagen ME, Minthon L, Winblad B, Blennow K (2009) CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA*, 302:385–393

[70] Fagan AM, Roe CM, Xiong C, Mintun MA, Morris JC, Holtzman DM (2007) Cerebrospinal fluid tau/beta-amyloid(42) ratio as a prediction of cognitive decline in nondemented older adults. *Arch Neurol*, 64:343–349

[71] Janelidze S, Mattsson N, Palmqvist S, Smith R, Beach TG, Serrano GE, Chai X, Proctor NK, Eichenlaub U, Zetterberg H, Blennow K, Reiman EM, Stomrud E, Dage JL, Hansson O (2020) Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia. *Nat Med*, 26:379–386

[72] Palmqvist S, Janelidze S, Stomrud E, Zetterberg H, Karl J, Zink K, Bittner T, Mattsson N, Eichenlaub U, Blennow K, Hansson O (2019) Performance of Fully Automated Plasma Assays as Screening Tests for Alzheimer Disease-Related β-Amyloid Status. *JAMA Neurol*, 76:1060–1069

[73] Kern S, Syrjanen JA, Blennow K, Zetterberg H, Skoog I, Waern M, Hagen CE, van Harten AC, Knopman DS, Jack CR, Petersen RC, Mielke MM (2019) Association of Cerebrospinal Fluid Neurofilament Light Protein With Risk of Mild Cognitive Im-

pairment Among Individuals Without Cognitive Impairment. *JAMA Neurol*, 76:187–193

[74] Tarawneh R, D'Angelo G, Crimmins D, Herries E, Griest T, Fagan AM, Zipfel GJ, Ladenson JH, Morris JC, Holtzman DM (2016) Diagnostic and Prognostic Utility of the Synaptic Marker Neurogranin in Alzheimer Disease. *JAMA Neurol*, 73:561–571

[75] Liu W, Lin H, He X, Chen L, Dai Y, Jia W, Xue X, Tao J, Chen L (2020) Neurogranin as a cognitive biomarker in cerebrospinal fluid and blood exosomes for Alzheimer's disease and mild cognitive impairment. *Transl Psychiatry*, 10:1–9

[76] De Wolf F, Ghanbari M, Licher S, McRae-McKee K, Gras L, Weverling GJ, Wermeling P, Sedaghat S, Ikram MK, Waziry R, Koudstaal W, Klap J, Kostense S, Hofman A, Anderson R, Goudsmit J, Ikram MA (2020) Plasma tau, neurofilament light chain and amyloid-β levels and risk of dementia; a population-based cohort study. *Brain*, 143:1220–1232

[77] Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, Haas J, Ruprecht K, Paul F, Stähler C, Lang CJG, Meder B, Bartfai T, Meese E, Keller A (2013) A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol*, 14:R78

[78] Cheng L, Doecke JD, Sharples RA, Villemagne VL, Fowler CJ, Rembach A, Martins RN, Rowe CC, Macaulay SL, Masters CL, Hill AF (2015) Prognostic serum miRNA biomarkers associated with Alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment. *Mol Psychiatry*, 20:1188–1196

[79] Vergallo A, Lista S, Zhao Y, Lemercier P, Teipel SJ, Potier MC, Habert MO, Dubois B, Lukiw WJ, Hampel H (2021) MiRNA-15b and miRNA-125b are associated with regional Aβ-PET and FDG-PET uptake in cognitively normal individuals with subjective memory complaints. *Transl Psychiatry*, 11:1–11

[80] Barthel H, Gertz HJ, Dresel S, Peters O, Bartenstein P, Buerger K, Hiemeyer F, Wittemer-Rump SM, Seibyl J, Reininger C, Sabri O, Florbetaben Study Group (2011) Cerebral amyloid-β PET with florbetaben (18F) in patients with Alzheimer's disease and healthy controls: a multicentre phase 2 diagnostic study. *Lancet Neurol*, 10:424–435

[81] Vandenberghe R, Van Laere K, Ivanoiu A, Salmon E, Bastin C, Triau E, Hasselbalch S, Law I, Andersen A, Korner A, Minthon L, Garraux G, Nelissen N, Bormans G, Buckley C, Owenius R, Thurfjell L, Farrar G, Brooks DJ (2010) 18F-flutemetamol amyloid imaging in Alzheimer disease and mild cognitive impairment: a phase 2 trial. *Ann Neurol*, 68:319–329

[82] Fleisher AS, Pontecorvo MJ, Devous MD, Lu M, Arora AK, Truocchio SP, Aldea P, Flitter M, Locascio T, Devine M, Siderowf A, Beach TG, Montine TJ, Serrano GE, Curtis C, Perrin A, Salloway S, Daniel M, Wellman C, Joshi AD, Irwin DJ, Lowe VJ, Seeley WW, Ikonomovic MD, Masdeu JC, Kennedy I, Harris T, Navitsky M, Southekal S, Mintun MA (2020) Positron Emission Tomography Imaging With [18F]flortaucipir and Postmortem Assessment of Alzheimer Disease Neuropathologic Changes. *JAMA Neurol*:e200528

[83] Swanson CJ, Zhang Y, Dhadda S, Wang J, Kaplow J, Lai RYK, Lannfelt L, Bradley H, Rabe M, Koyama A, Reyderman L, Berry DA, Berry S, Gordon R, Kramer LD, Cummings JL (2021) A randomized, double-blind, phase 2b proof-of-concept clinical trial in early Alzheimer's disease with lecanemab, an anti-Aβ protofibril antibody. *Alzheimers Res Ther*, 13:80

[84] Van Dyck CH, Nygaard HB, Chen K, Donohue MC, Raman R, Rissman RA, Brewer JB, Koeppe RA, Chow TW, Rafii MS, Gessert D, Choi J, Turner RS, Kaye JA, Gale SA, Reiman EM, Aisen PS, Strittmatter SM (2019) Effect of AZD0530 on Cerebral Metabolic Decline in Alzheimer Disease: A Randomized Clinical Trial. *JAMA Neurol*, 76:1219

[85] Congdon EE, Sigurdsson EM (2018) Tau-targeting therapies for Alzheimer disease. *Nat Rev Neurol*, 14:399–415

[86] Howard R, Zubko O, Bradley R, Harper E, Pank L, O'Brien J, Fox C, Tabet N, Livingston G, Bentham P, McShane R, Burns A, Ritchie C, Reeves S, Lovestone S, Ballard C, Noble W, Nilforooshan R, Wilcock G, Gray R, for the Minocycline in Alzheimer Disease Efficacy (MADE) Trialist Group (2020) Minocycline at 2 Different Dosages vs Placebo for Patients With Mild Alzheimer Disease: A Randomized Clinical Trial. *JAMA Neurol*, 77:164

[87] Hannestad J, Koborsi K, Klutzaritz V, Chao W, Ray R, Páez A, Jackson S, Lohr S, Cummings JL, Kay G, Nikolich K, Braithwaite S (2020) Safety and tolerability of GRF6019 in mild-to-moderate Alzheimer's disease dementia. *Alzheimers Dement (N Y)*, 6:e12115

[88] Pringsheim T, Jette N, Frolkis A, Steeves TDL (2014) The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov Disord*, 29:1583–1590

[89] Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkmann J, Schrag AE, Lang AE (2017) Parkinson disease. *Nat Rev Dis Primers*, 3:17013

[90] Antony PMA, Diederich NJ, Krüger R, Balling R (2013) The hallmarks of Parkinson's disease. *FEBS J*, 280:5981–5993

[91] Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, Obeso J, Marek K, Litvan I, Lang AE, Halliday G, Goetz CG, Gasser T, Dubois B, Chan P, Bloem BR, Adler CH, Deuschl G (2015) MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord*, 30:1591–1601

[92] King AE, Mintz J, Royall DR (2011) Meta-analysis of 123I-MIBG cardiac scintigraphy for the diagnosis of Lewy body-related disorders. *Mov Disord*, 26:1218–1224

[93] Wang J, Li Y, Huang Z, Wan W, Zhang Y, Wang C, Cheng X, Ye F, Liu K, Fei G, Zeng M, Jin L (2018) Neuromelanin-sensitive magnetic resonance imaging features of the substantia nigra and locus coeruleus in de novo Parkinson's disease and its phenotypes. *Eur J Neurol*, 25:949–e73

[94] Dayan E, Browner N (2017) Alterations in striato-thalamo-pallidal intrinsic functional connectivity as a prodrome of Parkinson's disease. *Neuroimage Clin*, 16:313–318

[95] Donadio V, Incensi A, Leta V, Giannoccaro MP, Scaglione C, Martinelli P, Capellari S, Avoni P, Baruzzi A, Liguori R (2014) Skin nerve α-synuclein deposits: a biomarker for idiopathic Parkinson disease. *Neurology*, 82:1362–1369

[96] Adler CH, Dugger BN, Hentz JG, Hinni ML, Lott DG, Driver-Dunckley E, Mehta S, Serrano G, Sue LI, Duffy A, Intorcia A, Filon J, Pullen J, Walker DG, Beach TG (2016) Peripheral Synucleinopathy in Early Parkinson's Disease: Submandibular Gland Needle Biopsy Findings. *Mov Disord*, 31:250–256

[97] Mirelman A, Bernad-Elazari H, Thaler A, Giladi-Yacobi E, Gurevich T, Gana-Weisz M, Saunders-Pullman R, Raymond D, Doan N, Bressman SB, Marder KS, Alcalay RN, Rao AK, Berg D, Brockmann K, Aasly J, Waro BJ, Tolosa E, Vilas D, Pont-Sunyer C, Orr-Urtreger A, Hausdorff JM, Giladi N (2016) Arm swing as a potential new prodromal marker of Parkinson's disease. *Mov Disord*, 31:1527–1534

[98] Swanson CR, Berlyand Y, Xie SX, Alcalay RN, Chahine LM, Chen-Plotkin AS (2015) Plasma apolipoprotein A1 associates with age at onset and motor severity in early Parkinson's disease patients. *Mov Disord*, 30:1648–1656

[99] Parnetti L, Gaetani L, Eusebi P, Paciotti S, Hansson O, El-Agnaf O, Mollenhauer B, Blennow K, Calabresi P (2019) CSF and blood biomarkers for Parkinson's disease. *Lancet Neurol*, 18:573–586

[100] Ravanidis S, Bougea A, Papagiannakis N, Maniati M, Koros C, Simitsi AM, Bozi M, Pachi I, Stamelou M, Paraskevas GP, Kapaki E, Moraitou M, Michelakakis H, Stefanis L, Doxakis E (2020) Circulating Brain-Enriched MicroRNAs for Detection and Discrimination of Idiopathic and Genetic Parkinson's Disease. *Mov Disord*, 35:457–467

[101] Khoo SK, Petillo D, Kang UJ, Resau JH, Berryhill B, Linder J, Forsgren L, Neuman LA, Tan AC (2012) Plasma-based circulating MicroRNA biomarkers for Parkinson's disease. *J Parkinsons Dis*, 2:321–331

[102] Dong H, Wang C, Lu S, Yu C, Huang L, Feng W, Xu H, Chen X, Zen K, Yan Q, Liu W, Zhang C, Zhang CY (2016) A panel of four decreased serum microRNAs as a novel biomarker for early Parkinson's disease. *Biomarkers*, 21:129–137

[103] Heinzel S, Berg D, Gasser T, Chen H, Yao C, Postuma RB, MDS Task Force on the Definition of Parkinson's Disease (2019) Update of the MDS research criteria for prodromal Parkinson's disease. *Mov Disord*, 34:1464–1470

[104] Frank MJ, Samanta J, Moustafa AA, Sherman SJ (2007) Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. *Science*, 318:1309–1312

[105] Hitti FL, Yang AI, Gonzalez-Alegre P, Baltuch GH (2019) Human gene therapy approaches for the treatment of Parkinson's disease: An overview of current and completed clinical trials. *Parkinsonism Relat Disord*, 66:16–24

[106] Schweitzer JS, Song B, Herrington TM, Park TY, Lee N, Ko S, Jeon J, Cha Y, Kim K, Li Q, Henchcliffe C, Kaplitt M, Neff C, Rapalino O, Seo H, Lee IH, Kim J, Kim T, Petsko GA, Ritz J, Cohen BM, Kong SW, Leblanc P, Carter BS, Kim KS (2020) Personalized iPSC-Derived Dopamine Progenitor Cells for Parkinson's Disease. *N Engl J Med*, 382:1926–1932

[107] Volc D, Poewe W, Kutzelnigg A, Lührs P, Thun-Hohenstein C, Schneeberger A, Galabova G, Majbour N, Vaikath N, El-Agnaf O, Winter D, Mihailovska E, Mairhofer A, Schwenke C, Staffler G, Medori R (2020) Safety and immunogenicity of the α-synuclein active immunotherapeutic PD01A in patients with Parkinson's disease: a randomised, single-blinded, phase 1 trial. *Lancet Neurol*, 19:591–600

[108] López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G (2013) The Hallmarks of Aging. *Cell*, 153:1194–1217

[109] Timmers PRHJ, Wilson JF, Joshi PK, Deelen J (2020) Multivariate genomic scan implicates novel loci and haem metabolism in human ageing. *Nat Commun*, 11:3570

[110] Lehallier B, Gate D, Schaum N, Nanasi T, Lee SE, Yousef H, Moran Losada P, Berdnik D, Keller A, Verghese J, Sathyan S, Franceschi C, Milman S, Barzilai N, Wyss-Coray T (2019) Undulating changes in human plasma proteome profiles across the lifespan. *Nat Med*, 25:1843–1850

[111] Schaum N, Lehallier B, Hahn O, Pálovics R, Hosseinzadeh S, Lee SE, Sit R, Lee DP, Losada PM, Zardeneta ME, Fehlmann T, Webber JT, McGeever A, Calcuttawala K, Zhang H, Berdnik D, Mathur V, Tan W, Zee A, Tan M, Tabula Muris Consortium, Pisco AO, Karkanias J, Neff NF, Keller A, Darmanis S, Quake SR, Wyss-Coray T (2020) Ageing hallmarks exhibit organ-specific temporal signatures. *Nature*, 583:596–602

[112] Tabula Muris Consortium (2020) A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583:590–595

[113] Horvath S, Raj K (2018) DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet*, 19:371–384

[114] Huan T, Chen G, Liu C, Bhattacharya A, Rong J, Chen BH, Seshadri S, Tanriverdi K, Freedman JE, Larson MG, Murabito JM, Levy D (2018) Age-associated microRNA expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell*, 17:e12687

[115] Hahn O, Drews LF, Nguyen A, Tatsuta T, Gkioni L, Hendrich O, Zhang Q, Langer T, Pletcher S, Wakelam MJO, Beyer A, Grönke S, Partridge L (2019) A nutritional memory effect counteracts benefits of dietary restriction in old mice. *Nat Metab*, 1:1059–1073

[116] Ahmed A, Tollefsbol T (2001) Telomeres and telomerase: basic science implications for aging. *J Am Geriatr Soc*, 49:1105–1109

[117] Sarkar TJ, Quarta M, Mukherjee S, Colville A, Paine P, Doan L, Tran CM, Chu CR, Horvath S, Qi LS, Bhutani N, Rando TA, Sebastiano V (2020) Transient non-integrative expression of nuclear reprogramming factors promotes multifaceted amelioration of aging in human cells. *Nat Commun*, 11:1545

[118] Villeda SA, Plambeck KE, Middeldorp J, Castellano JM, Mosher KI, Luo J, Smith LK, Bieri G, Lin K, Berdnik D, Wabl R, Udeochu J, Wheatley EG, Zou B, Simmons DA, Xie XS, Longo FM, Wyss-Coray T (2014) Young blood reverses age-related impairments in cognitive function and synaptic plasticity in mice. *Nat Med*, 20:659–663

[119] Li Y, Luo J, Zhou H, Liao JY, Ma LM, Chen YQ, Qu LH (2008) Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote Giardia lamblia. *Nucleic Acids Res*, 36:6048–6055

[120] Zhu L, Liu X, Pu W, Peng Y (2018) tRNA-derived small non-coding RNAs in human disease. *Cancer Lett*, 419:1–7

[121] Zhu L, Ow DW, Dong Z (2018) Transfer RNA-derived small RNAs in plants. *Sci China Life Sci*, 61:155–161

[122] Kumar P, Kuscu C, Dutta A (2016) Biogenesis and Function of Transfer RNA-Related Fragments (tRFs). *Trends Biochem Sci*, 41:679–689

[123] Noller HF, Green R, Heilek G, Hoffarth V, Hüttenhofer A, Joseph S, Lee I, Lieberman K, Mankin A, Merryman C (1995) Structure and function of ribosomal RNA. *Biochem Cell Biol*, 73:997–1009

[124] Hoagland MB, Stephenson ML, Scott JF, Hecht LI, Zamecnik PC (1958) A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem*, 231:241–257

[125] Bohnsack MT, Sloan KE (2018) Modifications in small nuclear RNAs and their roles in spliceosome assembly and function. *Biol Chem*, 399:1265–1276

[126] Dieci G, Preti M, Montanini B (2009) Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, 94:83–88

[127] Lerner MR, Boyle JA, Hardin JA, Steitz JA (1981) Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. *Science*, 211:400–402

[128] Christov CP, Gardiner TJ, Szüts D, Krude T (2006) Functional Requirement of Noncoding Y RNAs for Human Chromosomal DNA Replication. *Mol Cell Biol*, 26:6993–7004

[129] Statello L, Guo CJ, Chen LL, Huarte M (2021) Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*, 22:96–118

[130] Gebert LFR, MacRae IJ (2019) Regulation of microRNA function in animals. *Nat Rev Mol Cell Biol*, 20:21–37

[131] Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD (2019) PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet*, 20:89–108

[132] Darzacq X, Jády BE, Verheggen C, Kiss AM, Bertrand E, Kiss T (2002) Cajal body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylation guide RNAs. *EMBO J*, 21:2746–2756

[133] Dana H, Chalbatani GM, Mahmoodzadeh H, Karimloo R, Rezaiean O, Moradzadeh A, Mehmandoost N, Moazzen F, Mazraeh A, Marmari V, Ebrahimi M, Rashno MM, Abadi SJ, Gharagouzlo E (2017) Molecular Mechanisms and Biological Functions of siRNA. *Int J Biomed Sci*, 13:48–57

[134] Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, Rivas E, Eddy SR, Finn RD, Bateman A, Petrov AI (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res*, 49:D192–D200, D1

[135] Chan PP, Lowe TM (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res*, 44:D184–D189, Database issue

[136] Wang J, Zhang P, Lu Y, Li Y, Zheng Y, Kan Y, Chen R, He S (2019) piRBase: a comprehensive database of piRNA sequences. *Nucleic Acids Res*, 47:D175–D180, D1

[137] Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res*, 47:D155–D162, D1

[138] Fang S, Zhang L, Guo J, Niu Y, Wu Y, Li H, Zhao L, Li X, Teng X, Sun X, Sun L, Zhang MQ, Chen R, Zhao Y (2018) NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res*, 46:D308–D314, D1

[139] Zuo Y, Zhu L, Guo Z, Liu W, Zhang J, Zeng Z, Wu Q, Cheng J, Fu X, Jin Y, Zhao Y, Peng Y (2021) tsRBase: a comprehensive database for expression and function of tsRNAs in multiple species. *Nucleic Acids Res*, 49:D1038–D1045, D1

[140] RNAcentral Consortium (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res*, 49:D212–D220, D1

[141] Ecker JR, Davis RW (1986) Inhibition of gene expression in plant cells by expression of antisense RNA. *Proc Natl Acad Sci U S A*, 83:5372–5376

[142] Guo S, Kemphues KJ (1995) par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell*, 81:611–620

[143] Pal-Bhadra M, Bhadra U, Birchler JA (1997) Cosuppression in Drosophila: Gene Silencing of Alcohol dehydrogenase by white-Adh Transgenes Is Polycomb Dependent. *Cell*, 90:479–490

[144] Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391:806–811

[145] Misquitta L, Paterson BM (1999) Targeted disruption of gene function in Drosophila by RNA interference (RNA-i): A role for nautilus in embryonic somatic muscle formation. *Proc Natl Acad Sci U S A*, 96:1451–1456

[146] Sánchez Alvarado A, Newmark PA (1999) Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc Natl Acad Sci U S A*, 96:5049–5054

[147] Cogoni C, Macino G (1999) Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase. *Nature*, 399:166–169

[148] Lohmann JU, Endl I, Bosch TCG (1999) Silencing of Developmental Genes in Hydra. *Dev Biol*, 214:211–214

[149] Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411:494–498

[150] Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J (2000) Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature*, 408:325–330

[151] Kisielow M, Kleiner S, Nagasawa M, Faisal A, Nagamine Y (2002) Isoform-specific knockdown and expression of adaptor protein ShcA using small interfering RNA. *Biochem J*, 363:1–5, Pt 1

[152] North BJ, Marshall BL, Borra MT, Denu JM, Verdin E (2003) The human Sir2 ortholog, SIRT2, is an NAD+-dependent tubulin deacetylase. *Mol Cell*, 11:437–444

[153] Alekseev OM, Richardson RT, Alekseev O, O'Rand MG (2009) Analysis of gene expression profiles in HeLa cells in response to overexpression or siRNA-mediated depletion of NASP. *Reprod Biol Endocrinol*, 7:45

[154] Zheng S, Wang X, Weng YH, Jin X, Ji JL, Guo L, Hu B, Liu N, Cheng Q, Zhang J, Bai H, Yang T, Xia XH, Zhang HY, Gao S, Huang Y (2018) siRNA Knockdown of RRM2 Effectively Suppressed Pancreatic Tumor Growth Alone or Synergistically with Doxorubicin. *Mol Ther Nucleic Acids*, 12:805–816

[155] Suhr OB, Coelho T, Buades J, Pouget J, Conceicao I, Berk J, Schmidt H, Waddington-Cruz M, Campistol JM, Bettencourt BR, Vaishnaw A, Gollob J, Adams D (2015) Efficacy and safety of patisiran for familial amyloidotic polyneuropathy: a phase II multi-dose study. *Orphanet J Rare Dis*, 10:109

[156] Demirjian S, Ailawadi G, Polinsky M, Bitran D, Silberman S, Shernan SK, Burnier M, Hamilton M, Squiers E, Erlich S, Rothenstein D, Khan S, Chawla LS (2017) Safety and Tolerability Study of an Intravenously Administered Small Interfering Ribonucleic Acid (siRNA) Post On-Pump Cardiothoracic Surgery in Patients at Risk of Acute Kidney Injury. *Kidney Int Rep*, 2:836–843

[157] Benitez-Del-Castillo JM, Moreno-Montañés J, Jiménez-Alfaro I, Muñoz-Negrete FJ, Turman K, Palumaa K, Sádaba B, González MV, Ruz V, Vargas B, Pañeda C, Martínez T, Bleau AM, Jimenez AI (2016) Safety and Efficacy Clinical Trials for SYL1001, a Novel Short Interfering RNA for the Treatment of Dry Eye Disease. *Invest Ophthalmol Vis Sci*, 57:6447–6454

[158] Solano ECR, Kornbrust DJ, Beaudry A, Foy JWD, Schneider DJ, Thompson JD (2014) Toxicological and pharmacokinetic properties of QPI-1007, a chemically modified synthetic siRNA targeting caspase 2 mRNA, following intravitreal injection. *Nucleic Acid Ther*, 24:258–266

[159] Fitzgerald K, White S, Borodovsky A, Bettencourt BR, Strahs A, Clausen V, Wijngaard P, Horton JD, Taubel J, Brooks A, Fernando C, Kauffman RS, Kallend D, Vaishnaw A, Simon A (2017) A Highly Durable RNAi Therapeutic Inhibitor of PCSK9. *N Engl J Med*, 376:41–51

[160] Jackson AL, Linsley PS (2010) Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nat Rev Drug Discov*, 9:57–67

[161] Olejniczak M, Polak K, Galka-Marciniak P, Krzyzosiak WJ (2011) Recent advances in understanding of the immunological off-target effects of siRNA. *Curr Gene Ther*, 11:532–543

[162] Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854

[163] Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403:901–906

[164] Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408:86–89

[165] Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science*, 294:853–858

[166] Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294:858–862

[167] Lee RC, Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294:862–864

[168] Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19:92–105

[169] Kern F, Krammes L, Danz K, Diener C, Kehl T, Küchler O, Fehlmann T, Kahraman M, Rheinheimer S, Aparicio-Puerta E, Wagner S, Ludwig N, Backes C, Lenhof HP, von Briesen H, Hart M, Keller A, Meese E (2021) Validation of human microRNA target pathways enables evaluation of target prediction tools. *Nucleic Acids Res*, 49:127–144

[170] Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, Backes C, Wendschlag A, Giese N, Tjaden C, Ott K, Werner J, Hackert T, Ruprecht K, Huwer H, Huebers J, Jacobs G, Rosenstiel P, Dommisch H, Schaefer A, Müller-Quernheim J, Wullich B, Keck B, Graf N, Reichrath J, Vogel B, Nebel A, Jager SU, Staehler P, Amarantos I, Boisguerin V, Staehler C, Beier M, Scheffler M, Büchler MW, Wischhusen J, Haeusler SFM, Dietl J, Hofmann S, Lenhof HP, Schreiber S, Katus HA, Rottbauer W, Meder B, Hoheisel JD, Franke A, Meese E (2011) Toward the blood-borne miRNome of human diseases. *Nat Methods*, 8:841–843

[171] Hébert SS, Horré K, Nicolaï L, Bergmans B, Papadopoulou AS, Delacourte A, De Strooper B (2009) MicroRNA regulation of Alzheimer's Amyloid precursor protein expression. *Neurobiol Dis*, 33:422–428

[172] Kamal MA, Mushtaq G, Greig NH (2015) Current Update on Synopsis of miRNA Dysregulation in Neurological Disorders. *CNS Neurol Disord Drug Targets*, 14:492–501

[173] Palanichamy JK, Rao DS (2014) miRNA dysregulation in cancer: towards a mechanistic understanding. *Front Genet*, 5:54

[174] Griffiths-Jones S (2004) The microRNA Registry. *Nucleic Acids Res*, 32:D109–D111, suppl_1

[175] Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T (2003) A uniform system for microRNA annotation. *RNA*, 9:277–279

[176] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 36:D154–D158, suppl_1

[177] Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42:D68–D73, D1

[178] Okamura K, Phillips MD, Tyler DM, Duan H, Chou Yt, Lai EC (2008) The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nat Struct Mol Biol*, 15:354–363

[179] Yang JS, Phillips MD, Betel D, Mu P, Ventura A, Siepel AC, Chen KC, Lai EC (2011) Widespread regulatory activity of vertebrate microRNA* species. *RNA*, 17:312–326

[180] Marco A, Hui JHL, Ronshaugen M, Griffiths-Jones S (2010) Functional shifts in insect microRNA evolution. *Genome Biol Evol*, 2:686–696

[181] Griffiths-Jones S, Hui JHL, Marco A, Ronshaugen M (2011) MicroRNA evolution by arm switching. *EMBO Rep*, 12:172–177

[182] Van Peer G, Lefever S, Anckaert J, Beckers A, Rihani A, Van Goethem A, Volders PJ, Zeka F, Ongenaert M, Mestdagh P, Vandesompele J (2014) miRBase Tracker: keeping track of microRNA annotation changes. *Database (Oxford)*, 2014:bau080

[183] Backes C, Khaleeq QT, Meese E, Keller A (2016) miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res*, 44:W110–116, W1

[184] Backes C, Meder B, Hart M, Ludwig N, Leidinger P, Vogel B, Galata V, Roth P, Menegatti J, Grässer F, Ruprecht K, Kahraman M, Grossmann T, Haas J, Meese E, Keller A (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res*, 44:e53

[185] Brown M, Suryawanshi H, Hafner M, Farazi TA, Tuschl T (2013) Mammalian miRNA curation through next-generation sequencing. *Front Genet*, 4:145

[186] Hansen TB, Kjems J, Bramsen JB (2011) Enhancing miRNA annotation confidence in miRBase by continuous cross dataset analysis. *RNA Biol*, 8:378–383

[187] Meng Y, Shao C, Wang H, Chen M (2012) Are all the miRBase-registered microRNAs true? A structure- and expression-based re-examination in plants. *RNA Biol*, 9:249–253

[188] Wang X, Liu XS (2011) Systematic Curation of miRBase Annotation Using Integrated Small RNA High-Throughput Sequencing Data for C. elegans and Drosophila. *Front Genet*, 2:25

[189] Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E, Peterson KJ (2015) A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu Rev Genet*, 49:213–242

[190] Fromm B, Domanska D, Høye E, Ovchinnikov V, Kang W, Aparicio-Puerta E, Johansen M, Flatmark K, Mathelier A, Hovig E, Hackenberg M, Friedländer MR, Peterson KJ (2020) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res*, 48:D132–D141, D1

[191] Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y, Hatzimichael E, Kirino Y, Honda S, Lally M, Ramratnam B, Comstock CES, Knudsen KE, Gomella L, Spaeth GL, Hark L, Katz LJ, Witkiewicz A, Rostami A, Jimenez SA, Hollingsworth MA, Yeh JJ, Shaw CA, McKenzie SE, Bray P, Nelson PT, Zupo S, Van Roosbroeck K, Keating MJ, Calin GA, Yeo C, Jimbo M, Cozzitorto J, Brody JR, Delgrosso K, Mattick JS, Fortina P, Rigoutsos I (2015) Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc Natl Acad Sci U S A*, 112:E1106–1115

[192] Friedländer MR, Lizano E, Houben AJS, Bezdan D, Báñez-Coronel M, Kudla G, Mateu-Huertas E, Kagerbauer B, González J, Chen KC, LeProust EM, Martí E, Estivill X (2014) Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol*, 15:R57

[193] Jha A, Panzade G, Pandey R, Shankar R (2015) A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Res*, 43:8713–8724

[194] Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R (2016) The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res*, 44:D20–26, D1

[195] Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, 40:D54–56, Database issue

[196] Guo Z, Kuang Z, Wang Y, Zhao Y, Tao Y, Cheng C, Yang J, Lu X, Hao C, Wang T, Cao X, Wei J, Li L, Yang X (2020) PmiREN: a comprehensive encyclopedia of plant miRNAs. *Nucleic Acids Res*, 48:D1114–D1121, D1

[197] Da Fonseca BHR, Domingues DS, Paschoal AR (2019) mirtronDB: a mirtron knowledge base. *Bioinformatics*, 35:3873–3874

[198] Cuperus JT, Fahlgren N, Carrington JC (2011) Evolution and functional diversification of MIRNA genes. *Plant Cell*, 23:431–442

[199] Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. *Cell*, 136:669–687

[200] Chang SS, Zhang Z, Liu Y (2012) RNA interference pathways in fungi: mechanisms and functions. *Annu Rev Microbiol*, 66:305–323

[201] Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21:4663–4670

[202] Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10:1957–1966

[203] Mogilyansky E, Rigoutsos I (2013) The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death Differ*, 20:1603–1614

[204] Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432:231–235

[205] Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev*, 18:3016–3027

[206] Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, Kim VN (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425:415–419

[207] Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 17:3011–3016

[208] Bohnsack MT, Czaplinski K, Gorlich D (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, 10:185–191

[209] Bernstein E, Caudy AA, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409:363–366

[210] Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, Ha I, Baillie DL, Fire A, Ruvkun G, Mello CC (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106:23–34

[211] Hutvágner G, McLachlan J, Pasquinelli AE, Bálint E, Tuschl T, Zamore PD (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293:834–838

[212] Zhang H, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W (2004) Single processing center models for human Dicer and bacterial RNase III. *Cell*, 118:57–68

[213] Iwasaki S, Kobayashi M, Yoda M, Sakaguchi Y, Katsuma S, Suzuki T, Tomari Y (2010) Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Mol Cell*, 39:292–299

[214] Kawamata T, Tomari Y (2010) Making RISC. *Trends Biochem Sci*, 35:368–376

[215] Meijer HA, Smith EM, Bushell M (2014) Regulation of miRNA strand selection: follow the leader? *Biochem Soc Trans*, 42:1135–1140

[216] Ghildiyal M, Xu J, Seitz H, Weng Z, Zamore PD (2010) Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *RNA*, 16:43–56

[217] Frank F, Sonenberg N, Nagar B (2010) Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature*, 465:818–822

[218] Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115:199–208

[219] Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115:209–216

[220] Jonas S, Izaurralde E (2015) Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet*, 16:421–433

[221] Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA, Shin C, Baek D, Hsu SH, Ghoshal K, Villén J, Bartel DP (2014) mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol Cell*, 56:104–115

[222] Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466:835–840

[223] Ding L, Spencer A, Morita K, Han M (2005) The developmental timing regulator AIN-1 interacts with miRISCs and may target the argonaute protein ALG-1 to cytoplasmic P bodies in C. elegans. *Mol Cell*, 19:437–447

[224] Rehwinkel J, Behm-Ansmant I, Gatfield D, Izaurralde E (2005) A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. *RNA*, 11:1640–1647

[225] Behm-Ansmant I, Rehwinkel J, Doerks T, Stark A, Bork P, Izaurralde E (2006) mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev*, 20:1885–1898

[226] Chen CYA, Zheng D, Xia Z, Shyu AB (2009) Ago-TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps. *Nat Struct Mol Biol*, 16:1160–1166

[227] Braun JE, Huntzinger E, Fauser M, Izaurralde E (2011) GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. *Mol Cell*, 44:120–133

[228] Braun JE, Truffault V, Boland A, Huntzinger E, Chang CT, Haas G, Weichenrieder O, Coles M, Izaurralde E (2012) A direct interaction between DCP1 and XRN1 couples mRNA decapping to 5′ exonucleolytic degradation. *Nat Struct Mol Biol*, 19:1324–1331

[229] Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448:83–86

[230] Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev*, 22:2773–2785

[231] Ender C, Krek A, Friedländer MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G (2008) A human snoRNA with microRNA-like functions. *Mol Cell*, 32:519–528

[232] Cheloufi S, Dos Santos CO, Chong MMW, Hannon GJ (2010) A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, 465:584–589

[233] Yoda M, Cifuentes D, Izumi N, Sakaguchi Y, Suzuki T, Giraldez AJ, Tomari Y (2013) Poly(A)-specific ribonuclease mediates 3′-end trimming of Argonaute2-cleaved precursor microRNAs. *Cell Rep*, 5:715–726

[234] Lewis BP, Burge CB, Bartel DP (2005) Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120:15–20

[235] Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park SW, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WJ, Haussler D, Lai EC, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celniker SE, Gelbart WM, Kellis M (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450:219–232

[236] Schnall-Levin M, Zhao Y, Perrimon N, Berger B (2010) Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3′UTRs. *Proc Natl Acad Sci U S A*, 107:15751–15756

[237] Piwecka M, Glažar P, Hernandez-Miranda LR, Memczak S, Wolf SA, Rybak-Wolf A, Filipchyk A, Klironomos F, Cerda Jara CA, Fenske P, Trimbuch T, Zywitza V, Plass M, Schreyer L, Ayoub S, Kocks C, Kühn R, Rosenmund C, Birchmeier C, Rajewsky N (2017) Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science*, 357:eaam8526

[238] Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, 136:215–233

[239] Lewis BP, Shih Ih, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of Mammalian MicroRNA Targets. *Cell*, 115:787–798

[240] Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, 4

[241] Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27:91–105

[242] Wee LM, Flores-Jasso CF, Salomon WE, Zamore PD (2012) Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*, 151:1055–1067

[243] Kim D, Sung YM, Park J, Kim S, Kim J, Park J, Ha H, Bae JY, Kim S, Baek D (2016) General rules for functional microRNA targeting. *Nat Genet*, 48:1517–1526

[244] Moore MJ, Scheel TKH, Luna JM, Park CY, Fak JJ, Nishiuchi E, Rice CM, Darnell RB (2015) miRNA-target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nat Commun*, 6:8864

[245] Helwak A, Kudla G, Dudnakova T, Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153:654–665

[246] Hart M, Kern F, Backes C, Rheinheimer S, Fehlmann T, Keller A, Meese E (2018) The deterministic role of 5-mers in microRNA-gene targeting. *RNA Biol*, 15:819–825

[247] Bartel DP (2018) Metazoan MicroRNAs. *Cell*, 173:20–51

[248] Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, 18:610–621

[249] Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foà R, Schliwka J, Fuchs U, Novosel A, Müller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, De Vita G, Frezzetti D, Trompeter HI, Hornung V, Teng G, Hartmann G, Palkovits M, Di Lauro R, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, Tuschl T (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129:1401–1414

[250] Tan GC, Chan E, Molnar A, Sarkar R, Alexieva D, Isa IM, Robinson S, Zhang S, Ellis P, Langford CF, Guillot PV, Chandrashekran A, Fisk NM, Castellano L, Meister G, Winston RM, Cui W, Baulcombe D, Dibb NJ (2014) 5′ isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res*, 42:9424–9435

[251] Kim B, Jeong K, Kim VN (2017) Genome-wide Mapping of DROSHA Cleavage Sites on Primary MicroRNAs and Noncanonical Substrates. *Mol Cell*, 66:258–269.e5

[252] Nishikura K (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol*, 17:83–96

[253] Neilsen CT, Goodall GJ, Bracken CP (2012) IsomiRs–the overlooked repertoire in the dynamic microRNAome. *Trends Genet*, 28:544–549

[254] Telonis AG, Rigoutsos I (2018) Race Disparities in the Contribution of miRNA Isoforms and tRNA-Derived Fragments to Triple-Negative Breast Cancer. *Cancer Res*, 78:1140–1154

[255] Juzenas S, Venkatesh G, Hübenthal M, Hoeppner MP, Du ZG, Paulsen M, Rosenstiel P, Senger P, Hofmann-Apitius M, Keller A, Kupcinskas L, Franke A, Hemmrich-Stanisak G (2017) A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Res*, 45:9290–9301

[256] Fernandez-Valverde SL, Taft RJ, Mattick JS (2010) Dynamic isomiR regulation in Drosophila development. *RNA*, 16:1881–1888

[257] Guo L, Liang T, Yu J, Zou Q (2016) A Comprehensive Analysis of miRNA/isomiR Expression with Gender Difference. *PLoS One*, 11:e0154955

[258] Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res*, 45:2973–2985

[259] Magee RG, Telonis AG, Loher P, Londin E, Rigoutsos I (2018) Profiles of miRNA Isoforms and tRNA Fragments in Prostate Cancer. *Sci Rep*, 8:5314

[260] Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I (2015) Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Res*, 43:9158–9175

[261] Lee HY, Doudna JA (2012) TRBP alters human precursor microRNA processing in vitro. *RNA*, 18:2012–2019

[262] Fukunaga R, Han BW, Hung JH, Xu J, Weng Z, Zamore PD (2012) Dicer partner proteins tune the length of mature miRNAs in flies and mammals. *Cell*, 151:533–546

[263] Karali M, Persico M, Mutarelli M, Carissimo A, Pizzo M, Singh Marwah V, Ambrosio C, Pinelli M, Carrella D, Ferrari S, Ponzin D, Nigro V, di Bernardo D, Banfi S (2016) High-resolution analysis of the human retina miRNome reveals isomiR variations and novel microRNAs. *Nucleic Acids Res*, 44:1525–1540

[264] Manzano M, Forte E, Raja AN, Schipma MJ, Gottwein E (2015) Divergent target recognition by coexpressed 5′-isomiRs of miR-142-3p and selective viral mimicry. *RNA*, 21:1606–1620

[265] Llorens F, Bañez-Coronel M, Pantano L, del Río JA, Ferrer I, Estivill X, Martí E (2013) A highly expressed miR-101 isomiR is a functional silencing small RNA. *BMC Genomics*, 14:104

[266] Yu F, Pillman KA, Neilsen CT, Toubia J, Lawrence DM, Tsykin A, Gantier MP, Callen DF, Goodall GJ, Bracken CP (2017) Naturally existing isoforms of miR-222 have distinct functions. *Nucleic Acids Res*, 45:11371–11385

[267] Yamane D, Selitsky SR, Shimakami T, Li Y, Zhou M, Honda M, Sethupathy P, Lemon SM (2017) Differential hepatitis C virus RNA target site selection and host factor activities of naturally occurring miR-122 3 variants. *Nucleic Acids Res*, 45:4743–4755

[268] Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M (2011) Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res*, 21:1450–1461

[269] Menezes MR, Balzeau J, Hagan JP (2018) 3′ RNA Uridylation in Epitranscriptomics, Gene Regulation, and Disease. *Front Mol Biosci*, 5:61

[270] Kawahara Y, Megraw M, Kreider E, Iizasa H, Valente L, Hatzigeorgiou AG, Nishikura K (2008) Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res*, 36:5270–5280

[271] Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, Nishikura K (2006) Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol*, 13:13–21

[272] Kawahara Y, Zinshteyn B, Chendrimada TP, Shiekhattar R, Nishikura K (2007) RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. *EMBO Rep*, 8:763–769

[273] Vesely C, Tauber S, Sedlazeck FJ, Tajaddod M, von Haeseler A, Jantsch MF (2014) ADAR2 induces reproducible changes in sequence and abundance of mature microRNAs in the mouse brain. *Nucleic Acids Res*, 42:12155–12168

[274] Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science*, 315:1137–1140

[275] Nigita G, Acunzo M, Romano G, Veneziano D, Laganà A, Vitiello M, Wernicke D, Ferro A, Croce CM (2016) microRNA editing in seed region aligns with cellular changes in hypoxic conditions. *Nucleic Acids Res*, 44:6298–6308

[276] Schellenberg GD, D'Souza I, Poorkaj P (2000) The genetics of Alzheimer's disease. *Curr Psychiatry Rep*, 2:158–164

[277] Stracquadanio G, Wang X, Wallace MD, Grawenda AM, Zhang P, Hewitt J, Zeron-Medina J, Castro-Giner F, Tomlinson IP, Goding CR, Cygan KJ, Fairbrother WG, Thomas LF, Sætrom P, Gemignani F, Landi S, Schuster-Böckler B, Bell DA, Bond GL (2016) The importance of p53 pathway genetics in inherited and somatic cancer genomes. *Nat Rev Cancer*, 16:251–265

[278] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29:308–311

[279] Zhang Y, Fan M, Wang Q, He G, Fu Y, Li H, Yu S (2015) Polymorphisms in MicroRNA Genes And Genes Involving in NMDAR Signaling and Schizophrenia: A Case-Control Study in Chinese Han Population. *Sci Rep*, 5:12984

[280] Mencía A, Modamio-Høybjør S, Redshaw N, Morín M, Mayo-Merino F, Olavarrieta L, Aguirre LA, del Castillo I, Steel KP, Dalmay T, Moreno F, Moreno-Pelayo MA (2009) Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat Genet*, 41:609–613

[281] Shen J, Ambrosone CB, DiCioccio RA, Odunsi K, Lele SB, Zhao H (2008) A functional polymorphism in the miR-146a gene and age of familial breast/ovarian cancer diagnosis. *Carcinogenesis*, 29:1963–1966

[282] Stegeman S, Moya L, Selth LA, Spurdle AB, Clements JA, Batra J (2015) A genetic variant of MDM4 influences regulation by multiple microRNAs in prostate cancer. *Endocr Relat Cancer*, 22:265–276

[283] Wang M, Du M, Ma L, Chu H, Lv Q, Ye D, Guo J, Gu C, Xia G, Zhu Y, Ding Q, Yuan L, Fu G, Tong N, Qin C, Yin C, Xu J, Zhang Z (2016) A functional variant in TP63 at 3q28 associated with bladder cancer risk by creating an miR-140-5p binding site. *Int J Cancer*, 139:65–74

[284] Weaver DB, Anzola JM, Evans JD, Reid JG, Reese JT, Childs KL, Zdobnov EM, Samanta MP, Miller J, Elsik CG (2007) Computational and transcriptional evidence for microRNAs in the honey bee genome. *Genome Biol*, 8:R97

[285] Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*, 40:37–52

[286] Kim YK, Kim B, Kim VN (2016) Re-evaluation of the roles of DROSHA, Export in 5, and DICER in microRNA biogenesis. *Proc Natl Acad Sci U S A*, 113:E1881–1889

[287] Tokar T, Pastrello C, Rossos AEM, Abovsky M, Hauschild AC, Tsay M, Lu R, Jurisica I (2018) mirDIP 4.1—integrative database of human microRNA target predictions. *Nucleic Acids Res*, 46:D360–D370, Database issue

[288] Pinzón N, Li B, Martinez L, Sergeeva A, Presumey J, Apparailly F, Seitz H (2017) microRNA target prediction programs predict many false positives. *Genome Res*, 27:234–245

[289] Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G, Vergoulis T, Dalamagas T, Hatzigeorgiou AG (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res*, 46:D239–D245, Database issue

[290] Huang HY, Lin YCD, Li J, Huang KY, Shrestha S, Hong HC, Tang Y, Chen YG, Jin CN, Yu Y, Xu JT, Li YM, Cai XX, Zhou ZY, Chen XH, Pei YY, Hu L, Su JJ, Cui SD, Wang F, Xie YY, Ding SY, Luo MF, Chou CH, Chang NW, Chen KW, Cheng YH, Wan XH, Hsu WL, Lee TY, Wei FX, Huang HD (2020) miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res*, 48:D148–D154, D1

[291] Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature*, 460:479–486

[292] Clément T, Salone V, Rederstorff M (2015) Dual luciferase gene reporter assays to study miRNA function. *Methods Mol Biol*, 1296:187–198

[293] Ritchie W, Rasko JEJ, Flamant S (2013) MicroRNA target prediction and validation. *Adv Exp Med Biol*, 774:39–53

[294] On Chemical Safety IP. Biomarkers In Risk Assessment: Validity And Validation (EHC 222, 2001). 2001. URL: `http://www.inchem.org/documents/ehc/ehc/ehc222.htm#1.0` (visited on 06/23/2021)

[295] Weber JA, Baxter DH, Zhang S, Huang DY, Huang KH, Lee MJ, Galas DJ, Wang K (2010) The MicroRNA Spectrum in 12 Body Fluids. *Clin Chem*, 56:1733–1741

[296] Zubakov D, Boersma AWM, Choi Y, van Kuijk PF, Wiemer EAC, Kayser M (2010) MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation. *Int J Legal Med*, 124:217–226

[297] Hanson EK, Lubenow H, Ballantyne J (2009) Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. *Anal Biochem*, 387:303–314

[298] Machado MT, Navega S, Dias F, de Sousa MJC, Teixeira AL, Medeiros R (2015) microRNAs for peripheral blood fraction identification: Origin, pathways and forensic relevance. *Life Sci*, 143:98–104

[299] Montagne A, Barnes SR, Sweeney MD, Halliday MR, Sagare AP, Zhao Z, Toga AW, Jacobs RE, Liu CY, Amezcua L, Harrington MG, Chui HC, Law M, Zlokovic BV (2015) Blood-brain barrier breakdown in the aging human hippocampus. *Neuron*, 85:296–302

[300] Ma R, Jiang T, Kang X (2012) Circulating microRNAs in cancer: origin, function and application. *J Exp Clin Cancer Res*, 31:38

[301] Keller A, Backes C, Haas J, Leidinger P, Maetzler W, Deuschle C, Berg D, Ruschil C, Galata V, Ruprecht K, Stähler C, Würstle M, Sickert D, Gogol M, Meder B, Meese E (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement*, 12:565–576

[302] Leidinger P, Keller A, Meese E (2012) MicroRNAs – Important Molecules in Lung Cancer Research. *Front. Genet.*, 2

[303] Cosín-Tomás M, Antonell A, Lladó A, Alcolea D, Fortea J, Ezquerra M, Lleó A, Martí MJ, Pallàs M, Sanchez-Valle R, Molinuevo JL, Sanfeliu C, Kaliman P (2017) Plasma miR-34a-5p and miR-545-3p as Early Biomarkers of Alzheimer's Disease: Potential and Limitations. *Mol Neurobiol*, 54:5550–5562

[304] Fransquet PD, Ryan J (2018) Micro RNA as a potential blood-based epigenetic biomarker for Alzheimer's disease. *Clin Biochem*, 58:5–14

[305] Margis R, Margis R, Rieder CRM (2011) Identification of blood microRNAs associated to Parkinson's disease. *J Biotechnol*, 152:96–101

[306] Serafin A, Foco L, Zanigni S, Blankenburg H, Picard A, Zanon A, Giannini G, Pichler I, Facheris MF, Cortelli P, Pramstaller PP, Hicks AA, Domingues FS, Schwienbacher C (2015) Overexpression of blood microRNAs 103a, 30b, and 29a in L-dopa-treated patients with PD. *Neurology*, 84:645–653

[307] Cressatti M, Juwara L, Galindez JM, Velly AM, Nkurunziza ES, Marier S, Canie O, Gornistky M, Schipper HM (2020) Salivary microR-153 and microR-223 Levels as Potential Diagnostic Biomarkers of Idiopathic Parkinson's Disease. *Mov Disord*, 35:468–477

[308] Leidinger P, Keller A, Borries A, Huwer H, Rohling M, Huebers J, Lenhof HP, Meese E (2011) Specific peripheral miRNA profiles for distinguishing lung cancer from COPD. *Lung Cancer*, 74:41–47

[309] Keller A, Leidinger P, Borries A, Wendschlag A, Wucherpfennig F, Scheffler M, Huwer H, Lenhof HP, Meese E (2009) miRNAs in lung cancer - Studying complex fingerprints in patient's blood cells by microarray experiments. *BMC Cancer*, 9:353

[310] Leidinger P, Brefort T, Backes C, Krapp M, Galata V, Beier M, Kohlhaas J, Huwer H, Meese E, Keller A (2015) High-throughput qRT-PCR validation of blood microRNAs in non-small cell lung cancer. *Oncotarget*, 7:4611–4623

[311] Jiang M, Zhang P, Hu G, Xiao Z, Xu F, Zhong T, Huang F, Kuang H, Zhang W (2013) Relative expressions of miR-205-5p, miR-205-3p, and miR-21 in tissues and serum of non-small cell lung cancer patients. *Mol Cell Biochem*, 383:67–75

[312] Pan J, Zhou C, Zhao X, He J, Tian H, Shen W, Han Y, Chen J, Fang S, Meng X, Jin X, Gong Z (2018) A two-miRNA signature (miR-33a-5p and miR-128-3p) in whole blood as potential biomarker for early diagnosis of lung cancer. *Sci Rep*, 8:16699

[313] Keller A, Fehlmann T, Ludwig N, Kahraman M, Laufer T, Backes C, Vogelmeier C, Diener C, Biertz F, Herr C, Jörres RA, Lenhof HP, Meese E, Bals R, COSYCONET Study Group (2018) Genome-wide MicroRNA Expression Profiles in COPD: Early Predictors for Cancer Development. *Genomics Proteomics Bioinformatics*, 16:162–171

[314] Wang R, Xu J, Liu H, Zhao Z (2017) Peripheral leukocyte microRNAs as novel biomarkers for COPD. *Int J Chron Obstruct Pulmon Dis*, 12:1101–1112

[315] Akbas F, Coskunpinar E, Aynaci E, Oltulu YM, Yildiz P (2012) Analysis of serum micro-RNAs as potential biomarker in chronic obstructive pulmonary disease. *Exp Lung Res*, 38:286–294

[316] Soeda S, Ohyashiki JH, Ohtsuki K, Umezu T, Setoguchi Y, Ohyashiki K (2013) Clinical relevance of plasma miR-106b levels in patients with chronic obstructive pulmonary disease. *Int J Mol Med*, 31:533–539

[317] Kahraman M, Röske A, Laufer T, Fehlmann T, Backes C, Kern F, Kohlhaas J, Schrörs H, Saiz A, Zabler C, Ludwig N, Fasching PA, Strick R, Rübner M, Beckmann MW, Meese E, Keller A, Schrauder MG (2018) MicroRNA in diagnosis and therapy monitoring of early-stage triple-negative breast cancer. *Sci Rep*, 8:11584

[318] Inns J, James V (2015) Circulating microRNAs for the prediction of metastasis in breast cancer patients diagnosed with early stage disease. *Breast*, 24:364–369

[319] Kleivi Sahlberg K, Bottai G, Naume B, Burwinkel B, Calin GA, Børresen-Dale AL, Santarpia L (2015) A serum microRNA signature predicts tumor relapse and survival in triple-negative breast cancer patients. *Clin Cancer Res*, 21:1207–1214

[320] Heneghan HM, Miller N, Lowery AJ, Sweeney KJ, Newell J, Kerin MJ (2010) Circulating microRNAs as Novel Minimally Invasive Biomarkers for Breast Cancer. *Ann Surg*, 251:499–505

[321] Zhao H, Shen J, Medico L, Wang D, Ambrosone CB, Liu S (2010) A Pilot Study of Circulating miRNAs as Potential Biomarkers of Early Stage Breast Cancer. *PLoS One*, 5:e13735

[322] Vogel B, Keller A, Frese KS, Leidinger P, Sedaghat-Hamedani F, Kayvanpour E, Kloos W, Backe C, Thanaraj A, Brefort T, Beier M, Hardt S, Meese E, Katus HA, Meder B (2013) Multivariate miRNA signatures as biomarkers for non-ischaemic systolic heart failure. *Eur Heart J*, 34:2812–2822

[323] Fan KL, Zhang HF, Shen J, Zhang Q, Li XL (2013) Circulating microRNAs levels in Chinese heart failure patients caused by dilated cardiomyopathy. *Indian Heart J*, 65:12–16

[324] Akat KM, Moore-McGriff D, Morozov P, Brown M, Gogakos T, Correa Da Rosa J, Mihailovic A, Sauer M, Ji R, Ramarathnam A, Totary-Jain H, Williams Z, Tuschl T, Schulze PC (2014) Comparative RNA-sequencing analysis of myocardial and circulating small RNAs in human heart failure and their utility as biomarkers. *Proc Natl Acad Sci U S A*, 111:11151–11156

[325] Scrutinio D, Conserva F, Passantino A, Iacoviello M, Lagioia R, Gesualdo L (2017) Circulating microRNA-150-5p as a novel biomarker for advanced heart failure: A genome-wide prospective study. *J Heart Lung Transplant*, 36:616–624

[326] Wu T, Chen Y, Du Y, Tao J, Zhou Z, Yang Z (2018) Serum Exosomal MiR-92b-5p as a Potential Biomarker for Acute Heart Failure Caused by Dilated Cardiomyopathy. *Cell Physiol Biochem*, 46:1939–1950

[327] Hecksteden A, Leidinger P, Backes C, Rheinheimer S, Pfeiffer M, Ferrauti A, Kellmann M, Sedaghat F, Meder B, Meese E, Meyer T, Keller A (2016) miRNAs and sports: tracking training status and potentially confounding diagnoses. *J Transl Med*, 14:219

[328] Polakovičová M, Musil P, Laczo E, Hamar D, Kyselovič J (2016) Circulating MicroRNAs as Potential Biomarkers of Exercise Response. *Int J Mol Sci*, 17:1553

[329] Kern F, Ludwig N, Backes C, Maldener E, Fehlmann T, Suleymanov A, Meese E, Hecksteden A, Keller A, Meyer T (2019) Systematic Assessment of Blood-Borne MicroRNAs Highlights Molecular Profiles of Endurance Sport and Carbohydrate Uptake. *Cells*, 8:E1045

[330] Baggish AL, Hale A, Weiner RB, Lewis GD, Systrom D, Wang F, Wang TJ, Chan SY (2011) Dynamic regulation of circulating microRNA during acute exhaustive exercise and sustained aerobic exercise training. *J Physiol*, 589:3983–3994

[331] Banzet S, Chennaoui M, Girard O, Racinais S, Drogou C, Chalabi H, Koulmann N (2013) Changes in circulating microRNAs levels with exercise modality. *J Appl Physiol*, 115:1237–1244

[332] Kahraman M, Laufer T, Backes C, Schrörs H, Fehlmann T, Ludwig N, Kohlhaas J, Meese E, Wehler T, Bals R, Keller A (2017) Technical Stability and Biological Variability in MicroRNAs from Dried Blood Spots: A Lung Cancer Therapy-Monitoring Showcase. *Clin Chem*, 63:1476–1488

[333] Backes C, Leidinger P, Altmann G, Wuerstle M, Meder B, Galata V, Mueller SC, Sickert D, Stähler C, Meese E, Keller A (2015) Influence of next-generation sequencing and storage conditions on miRNA patterns generated from PAXgene blood. *Anal Chem*, 87:8910–8916

[334] Li L, Zhu D, Huang L, Zhang J, Bian Z, Chen X, Liu Y, Zhang CY, Zen K (2012) Argonaute 2 complexes selectively protect the circulating microRNAs in cell-secreted microvesicles. *PLoS One*, 7:e46957

[335] Vickers KC, Palmisano BT, Shoucri BM, Shamburek RD, Remaley AT (2011) MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat Cell Biol*, 13:423–433

[336] Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO (2007) Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat Cell Biol*, 9:654–659

[337] Saliminejad K, Khorram Khorshid HR, Ghaffari SH (2019) Why have microRNA biomarkers not been translated from bench to clinic? *Future Oncol*, 15:801–803

[338] Bonneau E, Neveu B, Kostantin E, Tsongalis G, De Guire V (2019) How close are miRNAs from clinical practice? A perspective on the diagnostic and therapeutic market. *EJIFCC*, 30:114–127

[339] Gustafson D, Tyryshkin K, Renwick N (2016) microRNA-guided diagnostics in clinical samples. *Best Pract Res Clin Endocrinol Metab*, 30:563–575

[340] Keller A, Leidinger P, Vogel B, Backes C, ElSharawy A, Galata V, Mueller SC, Marquart S, Schrauder MG, Strick R, Bauer A, Wischhusen J, Beier M, Kohlhaas J, Katus HA, Hoheisel J, Franke A, Meder B, Meese E (2014) miRNAs can be generally associated with human pathologies as exemplified for miR-144*. *BMC Med*, 12:224

[341] Jenike AE, Halushka MK (2021) miR-21: a non-specific biomarker of all maladies. *Biomarker Res*, 9:18

[342] Meder B, Backes C, Haas J, Leidinger P, Stähler C, Großmann T, Vogel B, Frese K, Giannitsis E, Katus HA, Meese E, Keller A (2014) Influence of the confounding factors age and sex on microRNA profiles from peripheral blood. *Clin Chem*, 60:1200–1208

[343] Dluzen DF, Noren Hooten N, Zhang Y, Kim Y, Glover FE, Tajuddin SM, Jacob KD, Zonderman AB, Evans MK (2016) Racial differences in microRNA and gene expression in hypertensive women. *Sci Rep*, 6:35815

[344] Metias SM, Lianidou E, Yousef GM (2009) MicroRNAs in clinical oncology: at the crossroads between promises and problems. *J Clin Pathol*, 62:771–776

[345] Farooqi AA, Fayyaz S, Shatynska-Mytsyk I, Javed Z, Jabeen S, Yaylim I, Gasparri ML, Panici PB (2016) Is miR-34a a Well-equipped Swordsman to Conquer Temple of Molecular Oncology? *Chem Biol Drug Des*, 87:321–334

[346] Hanna J, Hossain GS, Kocerha J (2019) The Potential for microRNA Therapeutics and Clinical Research. *Front. Genet.*, 10

[347] Zhang S, Cheng Z, Wang Y, Han T (2021) The Risks of miRNA Therapeutics: In a Drug Target Perspective. *Drug Des Devel Ther*, 15:721–733

[348] Haussecker D (2014) Current issues of RNAi therapeutics delivery and development. *J Control Release*, 195:49–54

[349] Ochoa S, Milam VT (2020) Modified Nucleic Acids: Expanding the Capabilities of Functional Oligonucleotides. *Molecules*, 25:E4659

[350] Burnett JC, Rossi JJ (2012) RNA-based therapeutics: current progress and future prospects. *Chem Biol*, 19:60–71

[351] Peer D, Lieberman J (2011) Special delivery: targeted therapy with small RNAs. *Gene Ther*, 18:1127–1133

[352] Gebert LFR, Rebhan MAE, Crivelli SEM, Denzler R, Stoffel M, Hall J (2014) Miravirsen (SPC3649) can inhibit the biogenesis of miR-122. *Nucleic Acids Res*, 42:609–621

[353] Moldovan L, Batte KE, Trgovcich J, Wisler J, Marsh CB, Piper M (2014) Methodological challenges in utilizing miRNAs as circulating biomarkers. *J Cell Mol Med*, 18:371–390

[354] Schmittgen TD, Lee EJ, Jiang J, Sarkar A, Yang L, Elton TS, Chen C (2008) Real-time PCR quantification of precursor and mature microRNA. *Methods*, 44:31–38

[355] Mestdagh P, Hartmann N, Baeriswyl L, Andreasen D, Bernard N, Chen C, Cheo D, D'Andrade P, DeMayo M, Dennis L, Derveaux S, Feng Y, Fulmer-Smentek S, Gerstmayer B, Gouffon J, Grimley C, Lader E, Lee KY, Luo S, Mouritzen P, Narayanan A, Patel S, Peiffer S, Rüberg S, Schroth G, Schuster D, Shaffer JM, Shelton EJ, Silveria S, Ulmanella U, Veeramachaneni V, Staedtler F, Peters T, Guettouche T, Wong L, Vandesompele J (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat Methods*, 11:809–815

[356] Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SP, Collins FS (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet*, 22:164–167

[357] Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, 23:41–46

[358] Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470

[359] Kraemer S, Vaught JD, Bock C, Gold L, Katilius E, Keeney TR, Kim N, Saccomano NA, Wilcox SK, Zichi D, Sanders GM (2011) From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteomic assay. *PLoS One*, 6:e26332

[360] Panse S, Dong L, Burian A, Carus R, Schutkowski M, Reimer U, Schneider-Mergener J (2004) Profiling of generic anti-phosphopeptide antibodies and kinases with peptide microarrays using radioactive and fluorescence-based assays. *Mol Divers*, 8:291–299

[361] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, 98:288–295

[362] Wang H, Ach RA, Curry B (2007) Direct and sensitive miRNA profiling from low-input total RNA. *RNA*, 13:151–159

[363] Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, Dotto GP, Harshman K (2009) Impact of normalization on miRNA microarray expression profiling. *RNA*, 15:493–501

[364] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). URL: `https://www.genome.gov/sequencingcostsdata` (visited on 07/13/2021)

[365] Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. *Clin Chem*, 55:641–658

[366] Modi A, Vai S, Caramelli D, Lari M (2021) The Illumina Sequencing Protocol and the NovaSeq 6000 System. Mengoni A, Bacci G, Fondi M, editors, *Bacterial Pangenomics: Methods and Protocols*, pages 15–42: Springer US, New York, NY

[367] Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta*, 1842:1932–1941

[368] Hahn O, Fehlmann T, Zhang H, Munson CN, Vest RT, Borcherding A, Liu S, Villarosa C, Drmanac S, Drmanac R, Keller A, Wyss-Coray T (2021) CoolMPS for robust sequencing of single-nuclear RNAs captured by droplet-based method. *Nucleic Acids Res*, 49:e11

[369] Heinicke F, Zhong X, Zucknick M, Breidenbach J, Sundaram AYM, T. Flåm S, Leithaug M, Dalland M, Farmer A, Henderson JM, Hussong MA, Moll P, Nguyen L, McNulty A, Shaffer JM, Shore S, Yip HK, Vitkovska J, Rayner S, Lie BA, Gilfillan GD (2019) Systematic assessment of commercially available low-input miRNA library preparation kits. *RNA Biol*, 17:75–86

[370] Hess JF, Kohl TA, Kotrová M, Rönsch K, Paprotka T, Mohr V, Hutzenlaub T, Brüggemann M, Zengerle R, Niemann S, Paust N (2020) Library preparation for next generation sequencing: A review of automation strategies. *Biotechnol Adv*, 41:107537

[371] Wright C, Rajpurohit A, Burke EE, Williams C, Collado-Torres L, Kimos M, Brandon NJ, Cross AJ, Jaffe AE, Weinberger DR, Shin JH (2019) Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. *BMC Genomics*, 20:513

[372] Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*, 9:72–74

[373] Giraldez MD, Spengler RM, Etheridge A, Godoy PM, Barczak AJ, Srinivasan S, De Hoff PL, Tanriverdi K, Courtright A, Lu S, Khoory J, Rubio R, Baxter D, Driedonks TAP, Buermans HPJ, Nolte-'t Hoen ENM, Jiang H, Wang K, Ghiran I, Wang YE, Van Keuren-Jensen K, Freedman JE, Woodruff PG, Laurent LC, Erle DJ, Galas DJ, Tewari M (2018) Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat Biotechnol*, 36:746–757

[374] Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res*, 39:e141

[375] Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454

[376] Allison DB, Brown AW, George BJ, Kaiser KA (2016) Reproducibility: A tragedy of errors. *Nature*, 530:27–29

[377] Abbott A (2019) The science institutions hiring integrity inspectors to vet their papers. *Nature*, 575:430–433

[378] Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J (2021) Sustainable data analysis with Snakemake. *F1000Res*, 10:33

[379] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C (2017) Nextflow enables reproducible computational workflows. *Nat Biotechnol*, 35:316–319

[380] Voss K, Auwera GVd, Gentry J (2017) Full-stack genomics pipelining with GATK4 + WDL + Cromwell. *F1000Res*, 6

[381] Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A (2020) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res*, 48:W395–W402, W1

[382] Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J, Bioconda Team (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*, 15:475–476

[383] Merkel D (2014) Docker: lightweight Linux containers for consistent development and deployment. *Linux J.*, 2014:2:2

[384] Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific containers for mobility of compute. *PLoS One*, 12:e0177459

[385] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44:D733–745, D1

[386] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Charkhchi M, Cummins C, Da Rin Fioretto L, Davidson C, Dodiya K, El Houdaigui B, Fatima R, Gall A, Garcia Giron C, Grego T, Guijarro-Clarke C, Haggerty L, Hemrom A, Hourlier T, Izuogu OG, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Gonzalez Martinez J, Marugán JC, Maurel T, McMahon AC, Mohanan S, Moore B, Muffato M, Oheh DN, Paraschas D, Parker A, Parton A, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Steed E, Szpak M, Szuba M, Taylor K, Thormann A, Threadgold G, Walts B, Winterbottom A, Chakiachvili M, Chaubal A, De Silva N, Flint B, Frankish A, Hunt SE, IIsley GR, Langridge N, Loveland JE, Martin FJ, Mudge JM, Morales J, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Cunningham F, Yates AD, Zerbino DR, Flicek P (2021) Ensembl 2021. *Nucleic Acids Res*, 49:D884–D891, D1

[387] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress

ML, Flicek P (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*, 47:D766–D773, D1

[388] Simoneau J, Dumontier S, Gosselin R, Scott MS (2021) Current RNA-seq methodology reporting limits reproducibility. *Brief Bioinform*, 22:140–145

[389] Kanduri C, Domanska D, Hovig E, Sandve GK (2017) Genome build information is an essential part of genomic track files. *Genome Biol*, 18:175

[390] Love MI, Soneson C, Hickey PF, Johnson LK, Pierce NT, Shepherd L, Morgan M, Patro R (2020) Tximeta: Reference sequence checksums for provenance identification in RNA-seq. *PLoS Comput Biol*, 16:e1007664

[391] Aparicio-Puerta E, Lebrón R, Rueda A, Gómez-Martín C, Giannoukakos S, Jaspez D, Medina JM, Zubkovic A, Jurak I, Fromm B, Marchal JA, Oliver J, Hackenberg M (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res*, 47:W530–W535, W1

[392] Lu Y, Baras AS, Halushka MK (2018) miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics*, 19:275

[393] Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34:i884–i890

[394] Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17:10–12

[395] Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30:2114–2120

[396] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10:R25

[397] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9:357–359

[398] Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760

[399] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29:15–21

[400] Maaten Lvd, Hinton G (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605

[401] McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.*, 3:861

[402] Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW (2018) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*

[403] Kobak D, Linderman GC (2021) Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol*, 39:156–157

[404] Li J, Bushel PR, Chu TM, Wolfinger RD (2009) Principal Variance Components Analysis: Estimating Batch Effects in Microarray Gene Expression Data, *Batch Effects and Noise in Microarray Experiments*, pages 141–154: John Wiley & Sons, Ltd

[405] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 43:e47

[406] Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140

[407] Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15:550

[408] Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B*, 57:289–300

[409] Altman N, Krzywinski M (2017) P values and the search for significance. *Nat Methods*, 14:3–4

[410] Jiang X, Du L, Wang L, Li J, Liu Y, Zheng G, Qu A, Zhang X, Pan H, Yang Y, Wang C (2015) Serum microRNA expression signatures identified from genome-wide microRNA profiling serve as novel noninvasive biomarkers for diagnosis and recurrence of bladder cancer. *Int J Cancer*, 136:854–862

[411] Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. Association for Computing Machinery, New York, NY, USA

[412] Rehman O, Zhuang H, Muhamed Ali A, Ibrahim A, Li Z (2019) Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach. *Cancers (Basel)*, 11:E431

[413] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320

[414] Hübenthal M, Hemmrich-Stanisak G, Degenhardt F, Szymczak S, Du Z, Elsharawy A, Keller A, Schreiber S, Franke A (2015) Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease. *PLoS One*, 10:e0140155

[415] Breiman L (2001) Random Forests. *Machine Learning*, 45:5–32

[416] Friedman JH (2001) Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29:1189–1232

[417] Elias KM, Fendler W, Stawiski K, Fiascone SJ, Vitonis AF, Berkowitz RS, Frendl G, Konstantinopoulos P, Crum CP, Kedzierska M, Cramer DW, Chowdhury D (2017) Diagnostic potential for a serum miRNA neural network for detection of ovarian cancer. *Elife*, 6:e28932

[418] Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517

[419] Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH (2018) Relief-based feature selection: Introduction and review. *J Biomed Inform*, 85:189–203

[420] Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46:389–422

[421] Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review. *Int. J. Patt. Recogn. Artif. Intell.*, 23:687–719

[422] Bonnet E, Wuyts J, Rouzé P, Van de Peer Y (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20:2911–2917

[423] Auyeung VC, Ulitsky I, McGeary SE, Bartel DP (2013) Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell*, 152:844–858

[424] Fang W, Bartel DP (2015) The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. *Mol Cell*, 60:131–145

[425] Alarcón CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF (2015) N6-methyladenosine marks primary microRNAs for processing. *Nature*, 519:482–485

[426] Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17:991–1008

[427] Lai EC, Tomancak P, Williams RW, Rubin GM (2003) Computational identification of Drosophila microRNA genes. *Genome Biol*, 4:R42

[428] Xue C, Li F, He T, Liu GP, Li Y, Zhang X (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310

[429] Batuwita R, Palade V (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 25:989–995

[430] Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res*, 35:W339–344, Web Server issue

[431] Zheng X, Fu X, Wang K, Wang M (2020) Deep neural networks for human microRNA precursor detection. *BMC Bioinformatics*, 21:17

[432] Rezoun ASM, Mehedi Hasan MA, Bin Aziz AZ (2020) Supervised Deep Learning Methods for Human pre-miRNA Identification. *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 1098–1101

[433] Jha A, Shankar R (2013) miReader: Discovering Novel miRNAs in Species without Sequenced Genome. *PLoS One*, 8:e66857

[434] Mapleson D, Moxon S, Dalmay T, Moulton V (2013) MirPlex: a tool for identifying miRNAs in high-throughput sRNA datasets without a genome. *J Exp Zool B Mol Dev Evol*, 320:47–56

[435] Vitsios DM, Kentepozidou E, Quintais L, Benito-Gutiérrez E, van Dongen S, Davis MP, Enright AJ (2017) Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Res*, 45:e177

[436] Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP (2002) Prediction of plant microRNA targets. *Cell*, 110:513–520

[437] Kern F, Backes C, Hirsch P, Fehlmann T, Hart M, Meese E, Keller A (2020) What's the target: understanding two decades of in silico microRNA-target prediction. *Brief Bioinform*, 21:1999–2010

[438] Min H, Yoon S (2010) Got target? Computational methods for microRNA target prediction and their extension. *Exp Mol Med*, 42:233–244

[439] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS (2003) MicroRNA targets in Drosophila. *Genome Biol*, 5:R1

[440] Pla A, Zhong X, Rayner S (2018) miRAW: A deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts. *PLoS Comput Biol*, 14:e1006185

[441] Marco A (2018) SeedVicious: Analysis of microRNA target and near-target sites. *PLoS One*, 13:e0195532

[442] Leclercq M, Diallo AB, Blanchette M (2017) Prediction of human miRNA target genes using computationally reconstructed ancestral mammalian sequences. *Nucleic Acids Res*, 45:556–566

[443] Bhattacharya A, Ziebarth J, Cui Y (2014) PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res*, 42, Database issue

[444] Liu CJ, Fu X, Xia M, Zhang Q, Gu Z, Guo AY (2021) miRNASNP-v3: a comprehensive database for SNPs and disease-related variations in miRNAs and miRNA targets. *Nucleic Acids Res*, 49:D1276–D1281, D1

[445] Liu B, Li J, Cairns MJ (2014) Identifying miRNAs, targets and functions. *Brief Bioinform*, 15:1–19

[446] Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD (2016) Bioinformatic tools for microRNA dissection. *Nucleic Acids Res*, 44:24–44

[447] Lai X, Wolkenhauer O, Vera J (2016) Understanding microRNA-mediated gene regulatory networks through mathematical modelling. *Nucleic Acids Res*, 44:6019–6035

[448] Sales G, Coppe A, Bisognin A, Biasiolo M, Bortoluzzi S, Romualdi C (2010) MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Res*, 38:W352–359, Web Server issue

[449] Huang GT, Athanassiou C, Benos PV (2011) mirConnX: condition-specific mRNA-microRNA network integrator. *Nucleic Acids Res*, 39:W416–423, Web Server issue

[450] Minchington TG, Griffiths-Jones S, Papalopulu N (2020) Dynamical gene regulatory networks are tuned by transcriptional autoregulation with microRNA feedback. *Sci Rep*, 10:12960

[451] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102:15545–15550

[452] Keller A, Backes C, Lenhof HP (2007) Computation of significance scores of unweighted Gene Set Enrichment Analyses. *BMC Bioinformatics*, 8:290

[453] Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD (2021) PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res*, 49:D394–D403, D1

[454] Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B (2019) WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*, 47:W199–W205, W1

[455] Gerstner N, Kehl T, Lenhof K, Müller A, Mayer C, Eckhart L, Grammes NL, Diener C, Hart M, Hahn O, Walter J, Wyss-Coray T, Meese E, Keller A, Lenhof HP (2020) GeneTrail 3: advanced high-throughput enrichment analysis. *Nucleic Acids Res*, 48:W515–W520, W1

[456] Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*, 49:D325–D334, D1

[457] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*, 49:D545–D551, D1

[458] Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P (2020) The reactome pathway knowledgebase. *Nucleic Acids Res*, 48:D498–D503, D1

[459] Godard P, van Eyll J (2015) Pathway analysis from lists of microRNAs: common pitfalls and alternative strategy. *Nucleic Acids Res*, 43:3490–3497

[460] Bleazard T, Lamb JA, Griffiths-Jones S (2015) Bias in microRNA functional enrichment analysis. *Bioinformatics*, 31:1592–1598

[461] Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res*, 37:D98–104, Database issue

[462] Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q (2019) HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res*, 47:D1013–D1017, D1

[463] Li J, Han X, Wan Y, Zhang S, Zhao Y, Fan R, Cui Q, Zhou Y (2018) TAM 2.0: tool for MicroRNA set analysis. *Nucleic Acids Res*, 46:W180–W185, W1

[464] Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, Dalamagas T, Hatzigeorgiou AG (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res*, 43:W460–466, W1

[465] UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 47:D506–D515, D1

[466] Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Ganesan S, Goodsell DS, Ghosh S, Green RK, Guranović V, Guzenko D, Hudson BP, Lawson CL, Liang Y, Lowe R, Namkoong H, Peisach E, Persikova I, Randle C, Rose A, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Tao YP, Voigt M, Westbrook JD, Young JY, Zardecki C, Zhuravleva M (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res*, 49:D437–D451, D1

[467] Tomczak K, Czerwińska P, Wiznerowicz M (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, 19:A68–77

[468] Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res*, 46:D794–D801, D1

[469] GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348:648–660

[470] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41:D991–995, Database issue

[471] Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res*, 36:W5–9, Web Server issue

[472] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, 59:307–321

[473] Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD (2018) HMMER web server: 2018 update. *Nucleic Acids Res*, 46:W200–W204, W1

[474] Usage Statistics and Market Share of JavaScript Libraries for Websites, July 2021. 2021. URL: https://w3techs.com/technologies/overview/javascript_library (visited on 07/23/2021)

[475] Kehl T, Backes C, Kern F, Fehlmann T, Ludwig N, Meese E, Lenhof HP, Keller A (2017) About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget*, 8:107167–107175

[476] Hart M, Rheinheimer S, Leidinger P, Backes C, Menegatti J, Fehlmann T, Grässer F, Keller A, Meese E (2016) Identification of miR-34a-target interactions by a combined network based and experimental approach. *Oncotarget*, 7:34288–34299

[477] Diener C, Hart M, Alansary D, Poth V, Walch-Rückheim B, Menegatti J, Grässer F, Fehlmann T, Rheinheimer S, Niemeyer BA, Lenhof HP, Keller A, Meese E (2018) Modulation of intracellular calcium signaling by microRNA-34a-5p. *Cell Death Dis*, 9:1008

[478] Kern F, Aparicio-Puerta E, Li Y, Fehlmann T, Kehl T, Wagner V, Ray K, Ludwig N, Lenhof HP, Meese E, Keller A (2021) miRTargetLink 2.0-interactive miRNA target gene and target pathway networks. *Nucleic Acids Res*, 49:W409–W416, W1

[479] Heinzelmann J, Arndt M, Pleyers R, Fehlmann T, Hoelters S, Zeuschner P, Vogt A, Pryalukhin A, Schaeffeler E, Bohle RM, Gajda M, Janssen M, Stoeckle M, Junker K (2019) 4-miRNA Score Predicts the Individual Metastatic Risk of Renal Cell Carcinoma Patients. *Ann Surg Oncol*, 26:3765–3773

[480] Keller A, Ludwig N, Fehlmann T, Kahraman M, Backes C, Kern F, Vogelmeier CF, Diener C, Fischer U, Biertz F, Herr C, Jörres RA, Lenhof HP, Bals R, Meese E (2019) Low miR-150-5p and miR-320b Expression Predicts Reduced Survival of COPD Patients. *Cells*, 8:E1162

[481] Abu-Halima M, Häusler S, Backes C, Fehlmann T, Staib C, Nestel S, Nazarenko I, Meese E, Keller A (2017) Micro-ribonucleic acids and extracellular vesicles repertoire in the spent culture media is altered in women undergoing In Vitro Fertilization. *Sci Rep*, 7:13525

[482] Keller A, Fehlmann T, Backes C, Kern F, Gislefoss R, Langseth H, Rounge TB, Ludwig N, Meese E (2020) Competitive learning suggests circulating miRNA profiles for cancers decades prior to diagnosis. *RNA Biol*, 17:1416–1426

[483] Kayvanpour E, Gi WT, Sedaghat-Hamedani F, Lehmann DH, Frese KS, Haas J, Tappu R, Samani OS, Nietsch R, Kahraman M, Fehlmann T, Müller-Hennessen M, Weis T, Giannitsis E, Niederdränk T, Keller A, Katus HA, Meder B (2021) microRNA neural networks improve diagnosis of acute coronary syndrome (ACS). *J Mol Cell Cardiol*, 151:155–162

[484] Ayoubian H, Ludwig N, Fehlmann T, Menegatti J, Gröger L, Anastasiadou E, Trivedi P, Keller A, Meese E, Grässer FA (2019) Epstein-Barr Virus Infection of Cell Lines Derived from Diffuse Large B-Cell Lymphomas Alters MicroRNA Loading of the Ago2 Complex. *J Virol*, 93:e01297–18

[485] Ludwig N, Becker M, Schumann T, Speer T, Fehlmann T, Keller A, Meese E (2017) Bias in recent miRBase annotations potentially associated with RNA quality issues. *Sci Rep*, 7:5162

[486] Schwarz EC, Backes C, Knörck A, Ludwig N, Leidinger P, Hoxha C, Schwär G, Grossmann T, Müller SC, Hart M, Haas J, Galata V, Müller I, Fehlmann T, Eichler H, Franke A, Meder B, Meese E, Hoth M, Keller A (2016) Deep characterization of blood cell miRNomes by NGS. *Cell Mol Life Sci*, 73:3169–3181

[487] Keller A, Kreis S, Leidinger P, Maixner F, Ludwig N, Backes C, Galata V, Guerriero G, Fehlmann T, Franke A, Meder B, Zink A, Meese E (2017) miRNAs in Ancient Tissue Specimens of the Tyrolean Iceman. *Mol Biol Evol*, 34:793–801

[488] Palmieri V, Backes C, Ludwig N, Fehlmann T, Kern F, Meese E, Keller A (2018) IMOTA: an interactive multi-omics tissue atlas for the analysis of human miRNA-target interactions. *Nucleic Acids Res*, 46:D770–D775, D1

[489] Backes C, Ludwig N, Leidinger P, Huwer H, Tenzer S, Fehlmann T, Franke A, Meese E, Lenhof HP, Keller A (2016) Paired proteomics, transcriptomics and miRNomics in non-small cell lung cancers: known and novel signaling cascades. *Oncotarget*, 7:71514–71525

[490] Schmartz GP, Kern F, Fehlmann T, Wagner V, Fromm B, Keller A (2021) Encyclopedia of tools for the analysis of miRNA isoforms. *Brief Bioinform*, 22:bbaa346

[491] Solomon J, Kern F, Fehlmann T, Meese E, Keller A (2020) HumiR: Web Services, Tools and Databases for Exploring Human microRNA Data. *Biomolecules*, 10:E1576

[492] Hücker SM, Fehlmann T, Werno C, Weidele K, Lüke F, Schlenska-Lange A, Klein CA, Keller A, Kirsch S (2021) Single-cell microRNA sequencing method comparison and application to cell lines and circulating lung tumor cells. *Nat Commun*, 12:4316

[493] Kern F, Fehlmann T, Keller A (2020) On the lifetime of bioinformatics web services. *Nucleic Acids Res*, 48:12523–12533

[494] Fehlmann T, Kern F, Hirsch P, Steinhaus R, Seelow D, Keller A (2021) Aviator: a web service for monitoring the availability of web services. *Nucleic Acids Res*, 49:W46–W51, W1

[495] Guimarães P, Keller A, Fehlmann T, Lammert F, Casper M (2020) Deep-learning based detection of gastric precancerous conditions. *Gut*, 69:4–6

[496] Guimarães P, Keller A, Fehlmann T, Lammert F, Casper M (2021) Deep-learning based detection of eosinophilic esophagitis. *Endoscopy*

[497] Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, Sarveswara R, Händler K, Pickkers P, Aziz NA, Ktena S, Tran F, Bitzer M, Ossowski S, Casadei N, Herr C, Petersheim D, Behrends U, Kern F, Fehlmann T, Schommers P, Lehmann C, Augustin M, Rybniker J, Altmüller J, Mishra N, Bernardes JP, Krämer B, Bonaguro L, Schulte-Schrepping J, De Domenico E, Siever C, Kraut M, Desai M, Monnet B, Saridaki M, Siegel CM, Drews A, Nuesch-Germano M, Theis H, Heyckendorf J, Schreiber S, Kim-Hellmuth S, COVID-19 Aachen Study (COVAS), Nattermann J, Skowasch D, Kurth I, Keller A, Bals R, Nürnberg P, Rieß O, Rosenstiel P, Netea MG, Theis F, Mukherjee S, Backes M, Aschenbrenner AC, Ulas T, Deutsche COVID-19 Omics Initiative (DeCOI), Breteler MMB, Giamarellos-Bourboulis EJ, Kox M, Becker M, Cheran S, Woodacre MS, Goh EL, Schultze JL (2021) Swarm Learning for decentralized and confidential clinical machine learning. *Nature*, 594:265–270

[498] Sedaghat-Hamedani F, Haas J, Zhu F, Geier C, Kayvanpour E, Liss M, Lai A, Frese K, Pribe-Wolferts R, Amr A, Li DT, Samani OS, Carstensen A, Bordalo DM, Müller M, Fischer C, Shao J, Wang J, Nie M, Yuan L, Haßfeld S, Schwartz C, Zhou M, Zhou Z, Shu Y, Wang M, Huang K, Zeng Q, Cheng L, Fehlmann T, Ehlermann P, Keller A, Dieterich C, Streckfuß-Bömeke K, Liao Y, Gotthardt M, Katus HA, Meder B (2017) Clinical genetics and outcome of left ventricular non-compaction cardiomyopathy. *Eur Heart J*, 38:3449–3460

[499] Fischer U, Backes C, Fehlmann T, Galata V, Keller A, Meese E (2019) Prospect and challenge of detecting dynamic gene copy number increases in stem cells by whole genome sequencing. *J Mol Med (Berl)*, 97:1099–1111

[500] Galata V, Fehlmann T, Backes C, Keller A (2019) PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res*, 47:D195–D202, D1

[501] Laczny CC, Kiefer C, Galata V, Fehlmann T, Backes C, Keller A (2017) BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Res*, 45:W171–W179, W1

[502] Amand J, Fehlmann T, Backes C, Keller A (2019) DynaVenn: web-based computation of the most significant overlap between ordered sets. *BMC Bioinformatics*, 20:743

[503] Scholz SS, Dillmann M, Flohr A, Backes C, Fehlmann T, Millenaar D, Ukena C, Böhm M, Keller A, Mahfoud F (2020) Contemporary scientometric analyses using a novel web application: the science performance evaluation (SciPE) approach. *Clin Res Cardiol*, 109:810–818

[504] Millenaar D, Fehlmann T, Scholz S, Pavlicek V, Flohr A, Dillmann M, Böhm M, Keller A, Mahfoud F, Ukena C (2020) Research in Atrial Fibrillation: A Scientometric Analysis Using the Novel Web Application SciPE. *JACC Clin Electrophysiol*, 6:1008–1018

[505] Grammes N, Millenaar D, Fehlmann T, Kern F, Böhm M, Mahfoud F, Keller A (2020) Research Output and International Cooperation Among Countries During the COVID-19 Pandemic: Scientometric Analysis. *J Med Internet Res*, 22:e24514

[506] Yang AC, Kern F, Losada PM, Agam MR, Maat CA, Schmartz GP, Fehlmann T, Stein JA, Schaum N, Lee DP, Calcuttawala K, Vest RT, Berdnik D, Lu N, Hahn O, Gate D, McNerney MW, Channappa D, Cobos I, Ludwig N, Schulz-Schaeffer WJ, Keller A, Wyss-Coray T (2021) Dysregulation of brain and choroid plexus cell types in severe COVID-19. *Nature*, 595:565–571

[507] Deogharia M, Majumder M (2018) Guide snoRNAs: Drivers or Passengers in Human Disease? *Biology (Basel)*, 8:E1

[508] Ren S, Lin P, Wang J, Yu H, Lv T, Sun L, Du G (2020) Circular RNAs: Promising Molecular Biomarkers of Human Aging-Related Diseases via Functioning as an miRNA Sponge. *Mol Ther Methods Clin Dev*, 18:215–229

[509] Shi J, Zhang Y, Tan D, Zhang X, Yan M, Zhang Y, Franklin R, Shahbazi M, Mackinlay K, Liu S, Kuhle B, James ER, Zhang L, Qu Y, Zhai Q, Zhao W, Zhao L, Zhou C, Gu W, Murn J, Guo J, Carrell DT, Wang Y, Chen X, Cairns BR, Yang XL, Schimmel P, Zernicka-Goetz M, Cheloufi S, Zhang Y, Zhou T, Chen Q (2021) PANDORA-seq expands the repertoire of regulatory small RNAs by overcoming RNA modifications. *Nat Cell Biol*, 23:424–436

[510] Pritchard CC, Kroh E, Wood B, Arroyo JD, Dougherty KJ, Miyaji MM, Tait JF, Tewari M (2012) Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. *Cancer Prev Res* (*Phila*), 5:492–497

[511] Sunderland N, Skroblin P, Barwari T, Huntley RP, Lu R, Joshi A, Lovering RC, Mayr M (2017) MicroRNA Biomarkers and Platelet Reactivity: The Clot Thickens. *Circ Res*, 120:418–435

[512] Jamali AA, Kusalik A, Wu FX (2020) MDIPA: a microRNA-drug interaction prediction approach based on non-negative matrix factorization. *Bioinformatics*, 36:5061–5067

[513] Rukov JL, Wilentzik R, Jaffe I, Vinther J, Shomron N (2014) Pharmaco-miR: linking microRNAs and drug effects. *Brief Bioinform*, 15:648–659

[514] Liu X, Wang S, Meng F, Wang J, Zhang Y, Dai E, Yu X, Li X, Jiang W (2013) SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics*, 29:409–411

[515] McGaughey G, Walters WP, Goldman B (2016) Understanding covariate shift in model performance. *F1000Res*, 5:597

[516] Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors (2008) Dataset Shift in Machine Learning. MIT Press, Cambridge, MA, USA

[517] Datlinger P, Rendeiro AF, Boenke T, Senekowitsch M, Krausgruber T, Barreca D, Bock C (2021) Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods*, 18:635–642

[518] Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, Gate RE, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*, 36:89–94

[519] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161:1187–1201

[520] Hashimshony T, Wagner F, Sher N, Yanai I (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*, 2:666–673

[521] Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, Martorell AJ, Ransohoff RM, Hafler BP, Bennett DA, Kellis M, Tsai LH (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, 570:332–337

[522] Agarwal D, Sandor C, Volpato V, Caffrey TM, Monzón-Sandoval J, Bowden R, Alegre-Abarrategui J, Wade-Martins R, Webber C (2020) A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat Commun*, 11:4183

[523] Smajić S, Prada-Medina CA, Landoulsi Z, Dietrich C, Jarazo J, Henck J, Balachandran S, Pachchek S, Morris CM, Antony P, Timmermann B, Sauer S, Schwamborn JC, May P, Grünewald A, Spielmann M (2020) Single-cell sequencing of the human midbrain reveals glial activation and a neuronal state specific to Parkinson's disease. *medRxiv*:2020.09.28.20202812

[524] Yang AC, Vest RT, Kern F, Lee DP, Maat CA, Losada PM, Chen MB, Agam M, Schaum N, Khoury N, Calcuttawala K, Pálovics R, Shin A, Wang EY, Luo J, Gate D, Siegenthaler JA, McNerney MW, Keller A, Wyss-Coray T (2021) A human brain vascular atlas reveals diverse cell mediators of Alzheimer's disease risk. *bioRxiv*:2021.04.26.441262

[525] Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, Lee JI, Suh YL, Ku BM, Eum HH, Choi S, Choi YL, Joung JG, Park WY, Jung HA, Sun JM, Lee SH, Ahn JS, Park K, Ahn MJ, Lee HO (2020) Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun*, 11:2285

[526] Nielsen MM, Pedersen JS (2021) miRNA activity inferred from single cell mRNA expression. *Sci Rep*, 11:9170

[527] Ma L, Huang Y, Zhu W, Zhou S, Zhou J, Zeng F, Liu X, Zhang Y, Yu J (2011) An integrated analysis of miRNA and mRNA expressions in non-small cell lung cancers. *PLoS One*, 6:e26502

[528] Kenny A, Jiménez-Mateos EM, Zea-Sevilla MA, Rábano A, Gili-Manzanaro P, Prehn JHM, Henshall DC, Ávila J, Engel T, Hernández F (2019) Proteins and microRNAs are differentially expressed in tear fluid from patients with Alzheimer's disease. *Sci Rep*, 9:15437

[529] Castro-Vega LJ, Letouzé E, Burnichon N, Buffet A, Disderot PH, Khalifa E, Loriot C, Elarouci N, Morin A, Menara M, Lepoutre-Lussey C, Badoual C, Sibony M, Dousset B, Libé R, Zinzindohoue F, Plouin PF, Bertherat J, Amar L, de Reyniès A, Favier J, Gimenez-Roqueplo AP (2015) Multi-omics analysis defines core genomic alterations in pheochromocytomas and paragangliomas. *Nat Commun*, 6:6044

[530] Chen S, Lake BB, Zhang K (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol*, 37:1452–1457

[531] Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*, 14:865–868

[532] Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, Thomson Z, Weiss MD, Li XJ, Savage AK, Green RR, Torgerson TR, Bumol TF, Graybuck LT, Skene PJ (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife*, 10:e63632

[533] Isakova A, Neff N, Quake SR (2020) Single cell profiling of total RNA using Smart-seq-total. *bioRxiv*:2020.06.02.131060

[534] Wang N, Zheng J, Chen Z, Liu Y, Dura B, Kwak M, Xavier-Ferrucio J, Lu YC, Zhang M, Roden C, Cheng J, Krause DS, Ding Y, Fan R, Lu J (2019) Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. *Nat Commun*, 10:95

[535] Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, Wlotzka W, von Haeseler A, Zuber J, Ameres SL (2017) Thiol-linked alkylation of RNA to assess expression dynamics. *Nat Methods*, 14:1198–1204

[536] Reichholf B, Herzog VA, Fasching N, Manzenreither RA, Sowemimo I, Ameres SL (2019) Time-Resolved Small RNA Sequencing Unravels the Molecular Principles of MicroRNA Homeostasis. *Mol Cell*, 75:756–768.e7

[537] Nunn JS, Tiller J, Fransquet P, Lacaze P (2019) Public Involvement in Global Genomics Research: A Scoping Review. *Front Public Health*, 7:79

[538] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*, 36:875–879

[539] Rautiainen M, Marschall T (2020) GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol*, 21:253

[540] Rakocevic G, Semenyuk V, Lee WP, Spencer J, Browning J, Johnson IJ, Arsenijevic V, Nadj J, Ghose K, Suciu MC, Ji SG, Demir G, Li L, Toptaş BÇ, Dolgoborodov A, Pollex B, Spulber I, Glotova I, Kómár P, Stachyra AL, Li Y, Popovic M, Källberg M, Jain A, Kural D (2019) Fast and accurate genomic analyses using genome graphs. *Nat Genet*, 51:354–362

Acknowledgement

First and foremost, I am extremely grateful to my supervisor Andreas Keller for giving me the opportunity to work on cutting edge research topics in his group, and his constant support and mentorship. His enthusiasm and scientific curiosity were an inspiration and always motivated me to push further. I am also deeply indebted to Christina Backes, whose unwavering guidance and extensive knowledge helped me to become a better scientist. Her encouragements and advice were essential for the completion of this thesis. Furthermore, I would like to express my sincere gratitude to Eckart Meese, whose scientific guidance led more than once to the successful completion of a project. I also wish to express my deepest appreciation to my committee for spending their time to evaluate my thesis and providing insightful comments.

I am immensely thankful to all my former and current colleagues at the Chair for Clinical Bioinformatics, especially Valentina, Mustafa, and Fabian, for many insightful discussions and invaluable assistance. Further, I gratefully acknowledge the help of our secretary, Sabine Lessel, who supported me with all administrative related tasks. Many thanks go also to all collaboration partners and coauthors I had the opportunity to work with. They proved to me that scientific work is never only the result of one single person's effort, but a collaborative effort involving the expertise of many people.

I would also like to thank my family for their unconditional backing. Especially helpful to me during this time were also my friends Raphaël and Marta. They allowed me to keep in mind that there is also a life beyond work. Finally, the accomplishment of this thesis would not have been possible without the love and support of my girlfriend Marie, who was always there to encourage me, no matter what.

Curriculum Vitae

Aus datenschutzrechtlichen Gründen wird der Lebenslauf in der elektronischen Fassung der Dissertation nicht veröffentlicht.