Saarland University
Graduate school of computer science

Integrative computational approaches for studying stem cell differentiation and complex diseases

Dissertation

zur Erlangung des Grades des
Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

von

Mohamed Hamed

Saarbrucken

August 2015

Datum des Kolloquiums	10th August 2015
Dekan der Fakultät 6	Univ.Prof. Dr. Markus Bläser
Mitglieder des Prüfungsausschusses:	
Vorsitzende:	Prof. Dr. Verena Wolf
Gutachter:	Prof. Dr. Volkhard Helms
	Dr. Marcel H. Schulz
Wissenschaftlicher Beisitzer:	Dr. Glenn Lawyer

Acknowledgment

All thanks and praises due to my lord first and last. Thanks to my God for making this possible and emphasizing more in me that indeed, there is no might nor power except in Allah.

First, I would like to thank Prof. Volkhard Helms, my PhD supervisor, for his support and guidance during my thesis. I did enjoy working with him and indeed he enriched my academic and professional skills by his fruitful scientific discussions, brilliant ideas, and unending motivations. I also thank him for such great research environment, teamwork spirit, and promoting excellent scientific atmosphere at his lab.

My gratitude and deep appreciation to all people involved in this research work for being very helpful, friendly, and open to collaboration. Namely, Dr. Martina Paulsen, Prof. Dr. Lutz von Mueller, Prof. Ulrich Nuebel, Dr. Daniel Patrick Nitsche, and Dr. Mohammad Irhimeh.

I am truly thankful to all my friends and colleagues in the Helms research group for the great work environment, the incredible lovely memories that will always be in mind, and their honest feedback and critical comments that helped me to improve my work. Special thanks to Christian Spaniol for the perfect teamwork, due date's commitments, and the interesting research achievements we have completed together. I really enjoyed working with him.

My biggest thanks go to my parents for their love, their prayer for me, and the unconditional support throughout all these years. Their constant support to me is invaluable and no words could express my feelings towards them. Loving thanks to my mom and my dad.

I owe my deepest gratitude to my wife for her steadfast moral and spiritual support as well as her understanding for my workload during the whole PhD period. Thank you my wife for your love and for being the best partner I could have ever wished for. I am also thankful to our kids (Youssef and Salma) for being the lovely part of our life and –sometimes- for correcting my German language.

Finally, I would like to acknowledge support from DAAD (Deutscher Akademischer Austauschdienst), funding by DFG through SFB 1027 and both the graduate school of computer science, and GradUS global in Saarland University.

Saarbrucken, June 2015

Abstract

The biological functions of the molecular components (genes, proteins, miRNAs, siRNAs,..etc) of biological cells and mutations/perturbations thereof are tightly connected with cellular malfunctions and disease pathways. Moreover, these molecular elements interact with each other forming a complex interwoven regulatory machinery that governs, on one hand, regular cellular pathways, and on the other hand, their dysregulation or malfunction in pathological processes. Therefore, revealing these critical molecular interactions in complex living systems is being considered as one of the major goals of current systems biology.

In this dissertation, we introduce practical computational approaches implemented as freely available software tools to integrate heterogeneous sources of large-scale genomic data and unravel the combinatorial regulatory interactions between different molecular elements. First, we present an automated GRN pipeline that constructs the genomic regulatory machinery of a cell from expression, sequencing, and annotation datasets through three modules implemented as separated software components (plugins) and hosted by our software framework Mebitoo that aims at automation of bioinformatics workflows. Then, we extended this pipeline to a general integrative network-based approach that involves also post-transcriptional interactions and reports the computational analysis of gene and miRNA transcriptomes, DNA methylome, and somatic mutations. This workflow enables users to identify putative disease drivers and novel targets for therapeutic treatment. Regarding the incorporation of somatic mutations with other genomic data sets, a stand-alone pipeline named "SnvDMiR" was implemented to explore possible genomic proximity relationships between somatic variants and both differentially methylated CpG sites as well as differentially expressed miRNAs. Along the same lines, but targeting the effects of genomic mutations, we developed an NGS pipeline and applied it to two groups of bacterial isolates (nasal and invasive) to investigate the phylogenetic positions of the recently emerged t504 clone (Spa-type t504) in the Saarland province of Germany and to better understand the infectivity mechanism of the invasive group. Motivated by all of this, we developed TFmiR as a freely available web server for deep and integrative downstream analysis of combinatorial regulatory interactions between TFs/genes and miRNAs that are involved in the pathogenesis of human diseases.

In the frame of this thesis, we employed these approaches to investigate the molecular mechanisms of cellular differentiation (namely hematopoiesis) as an example for biological processes and human breast cancer and diabetes as examples for complex diseases.

In summary, the work presented in this thesis has led to the development of interesting computational approaches that have been made available as non-commercial software toolkits. The provided topological and functional analyses of our approaches as validated on cellular differentiation and complex diseases promotes them as reliable systems biology tools for researchers across the life science communities.

Deutsche Zusammenfassung

Die Funktionsweise verschiedener molekularer Elemente (Gene, Proteine, Mutationen, miRNAs, siRNAs,... etc.) ist mit den darunterliegenden zellulären Fehlfunktionen als auch mit Krankheits-assoziierten zellulären Signalwegen verknüpft. Darüber hinaus interagieren diese molekularen Elemente auch miteinander und bilden eine komplexe ineinander verwobene regulatorische Maschinerie, die wiederum zelluläre Signalwege oder auch Krankheitsentwicklungen auf zellulärer Ebene beeinflusst. Aufgrund dessen ist heutzutage die Aufklärung dieser molekularen Interaktionen in komplexen lebenden Systemen eines der Hauptziele der Systembiologie.

In dieser Dissertation stellen wir rechnerbasierte Ansätze vor welche als Software frei verfügbar sind und die Integration von großen genomischen Datensätzen als auch eine damit verbundene Aufklärung der kombinatorischen Vielfalt dieser regulatorischen Interaktionen zwischen den verschiedenen molekularen Elementen, ermöglichten. Dafür entwickelten wir anfangs eine automatisierte GRN Pipeline, welche die regulatorische Maschinerie einer Zelle auf der Grundlage von Daten zur Genexpression, über Sequenzierung als auch Annotierung von Datensätzen konstruiert. Diese Pipeline wurde in drei separate Module aufgeteilt, die alle als Software plugins verfügbar sind, und in unser Framework Mebitoo, welches bioinformatische Arbeitsabläufe automatisiert, integriert sind. Daraufhin erweiterten wir unser bisheriges Framework um einem allgemeinen und integrativen Netzwerk-basierten Ansatz, welcher post-transkriptionelle Interaktionen berücksichtigt und die rechnerbasierte Analyse von Genen als auch miRNA Transkriptomen, dem DNA Methylom und somatischen Mutationen mit einbezieht. Unser Ziel war es, dabei vermeintliche Verursacher von Krankheitsbildern als auch neue Ziele für die therapeutische Behandlung von Krankheiten zu identifizieren. Für die Integration somatischer Mutationen wurde eine eigenständige Pipeline namens „SnvDMiR“ entwickelt, welche die Analyse von möglichen genomischen Nachbarschaftsbeziehungen zwischen somatischen Mutationen und differentiell methylierten CpG Positionen als auch differentiell exprimierten miRNAs, ermöglicht. Für die Analyse von somatischen Mutationen entwickelten wir zudem eine NGS Pipeline und wendeten diese auf zwei unterschiedliche Gruppen von bakteriellen Isolaten (nasale und invasive) an, um einerseits die phylogenetische Position des kürzlich im Saarland aufgekommenen Klons t504 (Spa-type t504) zu untersuchen, aber auch um den Mechanismus, der zu einer Infektion durch invasive Stämme führt, besser zu verstehen. All dies motivierte uns dazu TFmiR als frei verfügbare Web-Applikation zu entwickeln, welche eine tief gehende integrative Analyse von den kombinatorischen regulatorischen Interaktionen zwischen TFs/Genen und miRNAs ermöglicht, die an der Krankheitsentwicklung im Menschen beteiligt sind.

Die entwickelten Methoden wurden auf die zelluläre Differenzierung (Hämatopoese), als Beispiel für einen biologischen Prozess, als auch auf Brustkrebs und Diabetes, als Beispiele für komplexe Krankheiten, angewendet um deren molekulare Mechanismen zu untersuchen.

Zusammenfassend hat diese Arbeit zur Entwicklung von interessanten, rechnergestützten Methoden geführt, welche als nicht-kommerzielle Software publiziert wurden. Die Validierung unserer Methoden anhand von topologischen und funktionsbasierten Analysen sowohl in zellulärer Differenzierung als auch komplexen Krankheiten, machen diese zu verlässlichen systembiologischen Werkzeugen für Wissenschaftler aus den unterschiedlichsten Naturwissenschaftsbereichen.

“The soul is of the affair of my lord. And mankind has not been given of knowledge except a little”

“My Lord, Increase me in knowledge”

Quran ,Ch17, vers 85 , and Ch20, vers 114

Table of Contents

List of Figures

List of Tables

1. Introduction and biological background

Synopsis

In this chapter, we frame the problem of the multiple cellular factors affecting the regulatory machinery inside the cell on different molecular levels such as transcription, post-transcription, and post-translation and therefore the need for integrating information from heterogeneous sources of genomic data. We discuss the biology of the Gene Regulatory Network (GRN) and tackle its connection to biological processes as well as to the pathology of diseases. Finally, we give an overview of the structure of this thesis, the objectives of each chapter, and outline our contribution to the field.

1.1 Introduction

The ultimate goal of the genomic revolution and of modern systems biology is elucidating the genetic causes and drivers behind cellular processes, disease pathways, and phenotypic characteristics of organisms. This requires having a blueprint which states the different conditions in which genetic molecules, such as genes, proteins, and miRNAs, interact to make a complex living system [1]. In the past, these molecular associations have been reported at a rather slow pace. For example, it took more than a decade from the discovery of the well-known tumor suppressor gene p53 to conclude that it formed a regulatory feedback loop with its key regulator MDM2 [2]. Nowadays, advances in sequencing and expression technologies enable the generation of large high throughput data sets that allow for genome wide association studies. Indeed, this has made at least a part of this goal closer within reach, namely that of unraveling the underlying regulatory interactions between genes in a living system, or the so-called gene regulatory network (GRN). The identification of regulatory networks will help in identifying hundreds of genes that are responsible for most genetic diseases and that could serve as a starting point for new therapeutic intervention [3].

1.2 Gene Regulatory Networks (GRN)

Gene regulation is a general name for the control of gene expression levels and ultimately relates the specific quantity of a target gene product (protein) to the context of biological processes in cellular organisms.

In general, when a regulator protein binds to the regulatory sites of a gene, an mRNA transcript is produced that is in general translated into a specific protein or set of proteins. These proteins are either structural ones that colonize themselves at the cell membrane or enzymes that catalyze a certain reaction or regulators for the expression levels of other genes. The last groups of proteins generally bind to DNA and are known as transcription factors (TF), the main key players in the regulation machinery (Figure 1-1).

These proteins either activate or repress other genes by binding to their promoter regions and in this way initiate or inhibit the production of other proteins, and so on. Such multiple and concurrent cellular events lead to a complex and interwoven gene regulation machinery.

A gene regulation system consists of group of target genes, regulatory genomic regions (cis-regions), group of regulators, and their interactions. The regulators are often proteins (TFs) if regulation occurs at the transcription level or small molecules such as miRNAs and metabolites if regulation occurs at post-transcription or post-translation levels, respectively.

The cis-regions serve as aggregators of the effects of all transcription factors involved in gene regulation. Through protein-specific binding sites the cis-regions recruit and bring in proximity single TFs or groups of TFs (TF complex) having specific regulatory properties, with the purpose of inducing precisely when, where, and at what rate a gene is to be transcribed [1].

A gene regulatory network is typically represented as a graph in which the nodes are genes and the edges between nodes represent gene interactions through which the products of one gene affect those of another. These regulatory links can be inducting or activating (the arrowheads), where an increase in the expression of one leads to an increase in the other, or inhibitory/ repressing (the dull end), where an increase in one leading to a decrease in the other. A series of edges indicates a chain of such dependencies, with cycles corresponding to feedback loops [4].

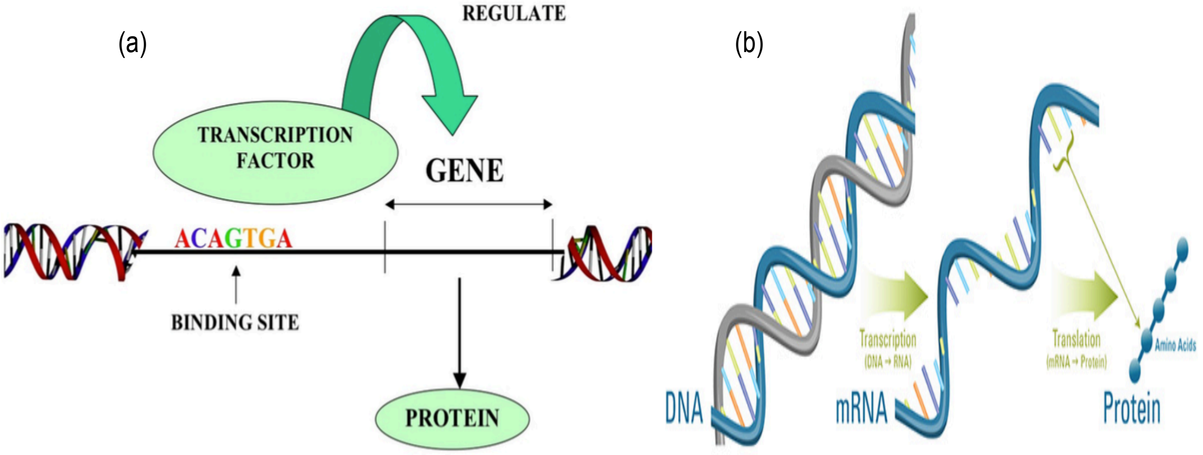


Figure 1-1 Regulation of gene transcription.
Schematic diagram illustrating how a transcription factor binds to the DNA at specific binding motifs in the promoter region of a gene, and thereby regulates the activities (rate of transcription) of this gene. (b) Transcription of genes into mRNA and translation of mRNA into amino acid chains (proteins). A cell's DNA carries the instructions, or genes, to make the proteins that are needed to build cell structures and to perform necessary functions. To make a protein, the instructions in the DNA are transcribed, or copied, to a molecule of messenger RNA (mRNA). Other molecules in the cell then help translating those instructions to assemble the protein by stringing together more than 20 different kinds of amino acids in a specific sequence. Messenger RNA provides vital clues about the processes a cell uses to survive, because it shows which genes are being used at a given time. Source: https://sbi4u2013.wordpress.com/author/viceteacher/, and http://www.whoi.edu/news-release/DeepBiosphere_mRNA.

1.3 Biological properties of GRN

Uncovering the architecture, dynamics, and the interwoven nature of the regulatory machinery in biological cells depends on our knowledge of the biological properties of gene networks. Noticeably more is known about the gene regulation circuitry today than few years ago, which helped scientists to effectively model GRNs and powerfully understand the underlying and controlled cellular behaviors of specific processes.

For example, one of the important properties of gene network topology (structure), which defines the connections between nodes, is their sparseness. This means that each gene is regulated only by relatively few other genes and consecutively; there is a small number of edges per node, smaller than the total number of nodes [5]. The sparseness property is often used to prune the search space and reduce the data dimensionality during network inference using data portioning methods (clustering and biclustering algorithms), as described later in chapter 3 of this thesis. It has been previously shown that the degree centrality distribution of biological networks tends to be longer tailed than the normal distribution [6]. The appropriate distribution seems to belong to the so-called scale-free networks. This is a class of networks where the frequency *P(N)* of nodes with *N* connections in the GRN graph (i.e. the degree of the node) depends on *N* by a power-law $P(N) = N^{-\gamma}$, where γ is some network specific constant. Such scale-free networks exhibit one important characteristic that is the emergence of hubs, or highly connected nodes in the network. Such hub nodes are extremely unlikely to happen in standard random graphs. . These hub nodes correspond to highly central nodes in the gene network, i.e. genes that contribute a large amount of the overall regulation [1]. They could in fact be potential candidates for master regulatory genes or essential genetic determinants of cell fate or probable targets for new drugs and treatment of complex diseases, as is demonstrated in chapter 6 and also in [7].

Another important feature of GRNs is the network modularity. GRNS are often composed of inhomogeneous and different kinds of subcircuits or modules that each have a specific kind of cellular function [8]. This concept is important, because it plays a key role for designing gene networks in synthetic biology which aims at designing novel biological circuits able to perform specific tasks (for example, the periodic expression of a gene of interest) [3].

1.4 Complexity of GRNs

The complexity of gene regulatory systems goes back to the different cellular levels (transcription, post-transcription, and post-translation) at which they can be modeled as well as the huge number of genetic molecules (genes, proteins, metabolites, miRNAs…) involved at each level. A widely used approach is to reduce or subsume the regulatory system to the gene space at the transcription level for the sake of simplicity and research feasibility (Figure 1-2). This also depends on the existing biological knowledge and the availability of empirical data, as well as on the goal of the project, which can be as simple as hypothesis testing, or as complex as quantitative network modeling. Although modeling the GRN on the gene space is a common approach nowadays, it remains insufficient to puzzle the complete picture of regulatory mechanisms because other important genetic factors that affect the regulation system at other levels are ignored. In turn, this will not help in fully elucidating the associated biological processes and functions of genetic molecules. To this end, tackling this problem was the spirit behind the work in this thesis.

1.5 Levels of gene regulation

There are three main levels of controlling gene expression in living cells as shown in Figure 1-3. We summarize them as follows:

1.5.1 Transcriptional regulation

Transcriptional regulation is the most common type of regulation. It regulates which genes are transcribed (from DNA to mRNA) and controls the rate of transcription or levels of gene expression. Transcription regulation includes two main cellular events. Firstly, the binding of a regulator molecule to the cis-regulatory region of a target gene to initiate the transcription process as described above in section 1-2. Secondly and sole importantly, are epigenetic modifications.

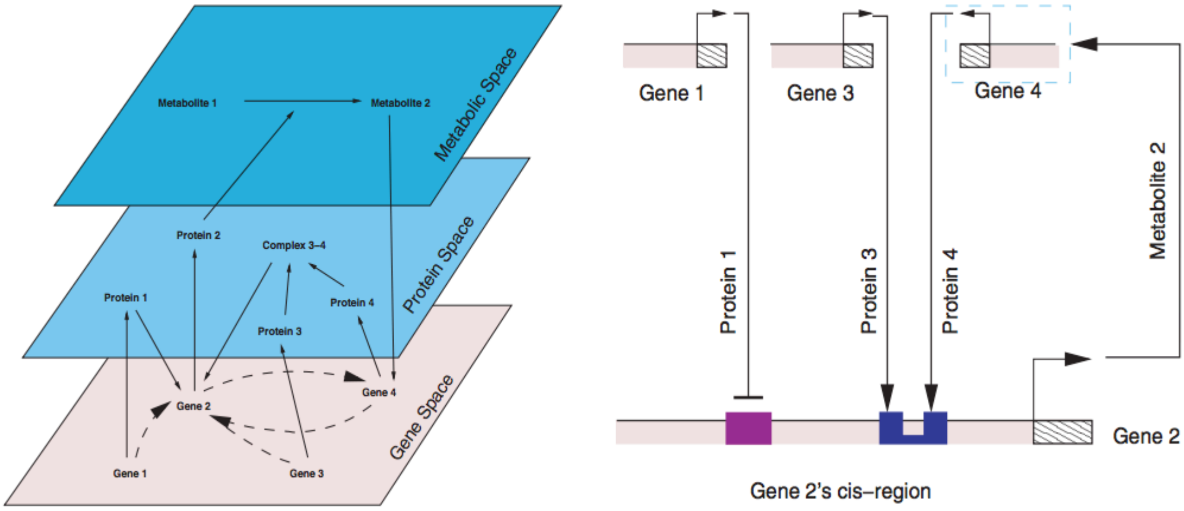


Figure 1-2 Complexity of gene regulation machinery and reducing it into gene space.
Shown on the left are the multiple levels at which genes are regulated by other genes, proteins and metabolites. On the right is a useful abstraction subsuming all the interactions into ones between genes only. The cis-regions are shown next to the coding regions, which are marked with pattern fill and start at the bent arrows. The edges are marked with the name of the molecule that carries the interaction. Some reactions represent transcription factor – DNA binding, happen during transcription, and are localized on the cis-regions. In those cases the corresponding protein-specific binding sites, or cis-elements, on the cis-regions are shown (colored polygons). Otherwise, the interactions can take place during transcription or later (e.g. post-translational modifications) as may be the case with Metabolite 2 interacting with Gene 4. The nature of the interactions is inducing (arrow) or repressing (dull end). Source: modified from [1].

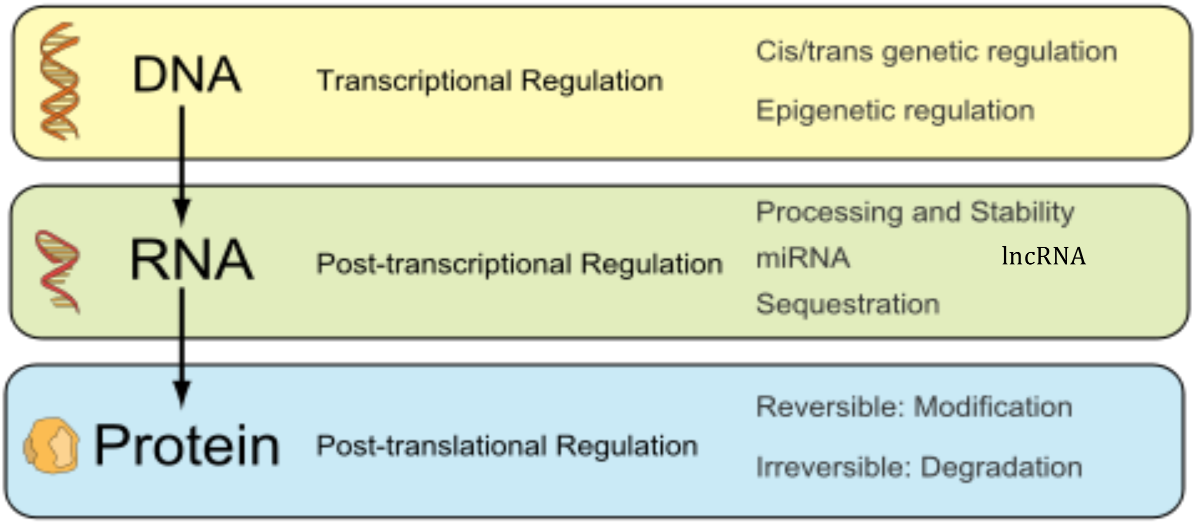

DNA
Transcriptional Regulation
Cis/trans genetic regulation
Epigenetic regulation
RNA
Post-transcriptional Regulation
Processing and Stability
miRNA
lncRNA
Sequestration
Protein
Post-translational Regulation
Reversible: Modification
Irreversible: Degradation

Figure 1-3 Different levels of gene regulation system.
Source: modified from http://en.wikipedia.org/wiki/File:Gene_Regulation.svg

1.5.1.1 Epigenetic modifications

The term "epigenetics" refers to the study of the heritable alterations and modifications in phenotypic expression that don't involve changes in DNA sequence [9]. These modifications were found to be highly correlated with the changes in DNA sequence through evolution [10]. Furthermore, these epigenetic changes are essential for normal development, biological cellular processes such as cell differentiation, and are increasingly recognized as being involved in genetic disorders or complex diseases like cancer [11]. Epigenetic regulations can switch genes on or off and determine which proteins are transcribed by specific genetic events other than an individual's DNA sequence.

Figure 1-4 illustrates the epigenetic modifications of the genome of an organism. These modifications include DNA methylation, histone modifications, and effects induced by non-coding RNAs. The regulation effects of non-coding RNAs will be discussed in the next section on post-transcriptional regulation events.

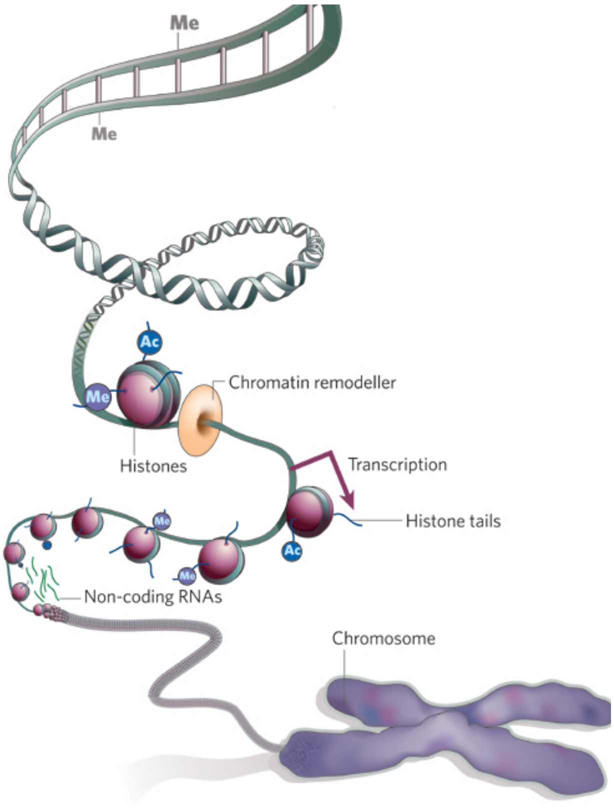


Figure 1-4 Epigenetic modifications. The genome is prone to direct methylation of DNA and histone modifications; which include histone acetylation and methylation. Other chromatin remodelers also come into play. Additionally, noncoding RNAs play a major role in DNA targeting by silencing or different mechanisms. Source: modified from [12].

1.5.1.1.1 DNA methylation

DNA methylation is the most extensively studied epigenetic modification that is being increasingly recognized to play an important role in the regulation of gene expression and is used as epigenetic marker for different disease pathways [13-16]. In mammals, DNA methylation typically occurs in a CpG dinucleotide context that is often grouped in clusters called CpG islands. More than half of the gene promoters in human are associated with CpG regions and are usually unmethylated in normal cells, although some of them become methylated in a tissue-specific manner during early development [17]. DNA methylation is also believed to be a crucial reason behind genomic imprinting (see next section), where hypermethylation at one of the two parental alleles leads to monoallelic expression [17]. DNA methylation profiling unravels differentially methylated regions (DMRs) that are in principle CpG sites altered during disease or oncogenic processes [18, 19] as shown in Figure 1-5. Hypermethylation of CpG islands located in promoter regions, for example, is involved in gene silencing at the transcriptional level [20] (Figure 1-6) and often leads to a high rate of C to T mutations at these sites [21].

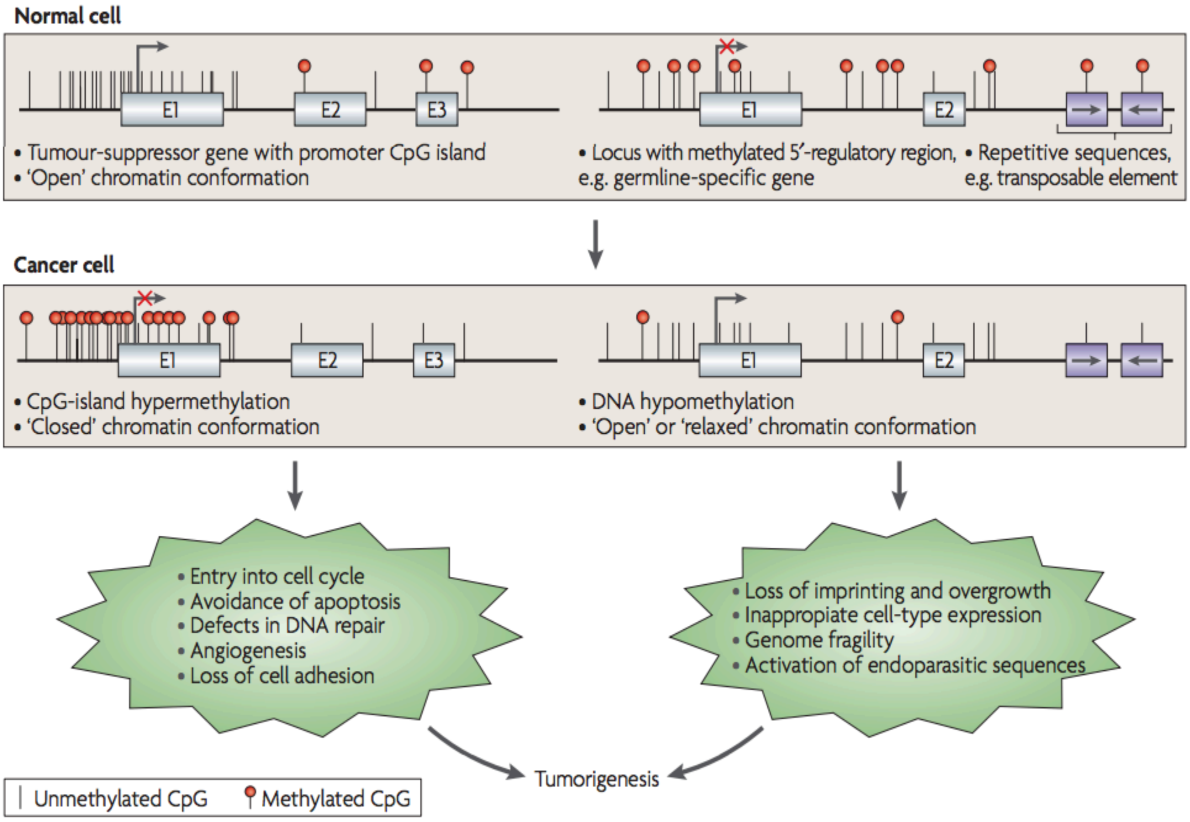

Normal cell
E1
E2
E3
• Tumour-suppressor gene with promoter CpG island
• 'Open' chromatin conformation
E1
E2
• Locus with methylated 5′-regulatory region, e.g. germline-specific gene
• Repetitive sequences, e.g. transposable element
Cancer cell
E1
E2
E3
• CpG-island hypermethylation
• 'Closed' chromatin conformation
E1
E2
• DNA hypomethylation
• 'Open' or 'relaxed' chromatin conformation
• Entry into cell cycle
• Avoidance of apoptosis
• Defects in DNA repair
• Angiogenesis
• Loss of cell adhesion
• Loss of imprinting and overgrowth
• Inappropiate cell-type expression
• Genome fragility
• Activation of endoparasitic sequences
Tumorigenesis
Unmethylated CpG
Methylated CpG

Figure 1-5 Altered DNA-methylation patterns in tumorigenesis.
The hypermethylation of CpG islands of tumor suppressor genes is a common alteration in cancer cells, and leads to the transcriptional inactivation of these genes and the loss of their normal cellular functions. This contributes to many of the hallmarks of cancer cells. At the same time, the genome of the cancer cell undergoes global hypomethylation at repetitive sequences, and tissue-specific and imprinted genes can also show loss of DNA methylation. In some cases, this hypomethylation is known to contribute to cancer cell phenotypes, causing changes such as loss of imprinting, and might also contribute to the genomic instability that characterizes tumors. E, exon. Source: modified from [19].

Genomic imprinting

One of the important epigenetic phenomena in mammals is genomic imprinting, by which certain genes are expressed in a parent-of-origin-specific manner. If the allele inherited from the father is imprinted, then it is silenced and the gene is called maternally expressed, and vice versa [22]. To date, about 100 genes have been experimentally confirmed to be imprinted in mammals. Thus, the imprinting phenomenon affects a fairly small number of genes. Many studies showed that imprinted genes are not only important during embryonic development but possess also postnatal functions. Hence, the kinship theory with its focus on prenatal development might explain some but not all aspects of the evolution of genomic imprinting.

During postnatal development, genomic imprinting affects endocrinal networks, energy metabolism, and behavior. Prominent examples for the functions of imprinted genes in endocrinal pathways are the imprinted transcripts of the *Gnas* locus. In the human, genetic and epigenetic aberrations in this region are associated with Albright hereditary osteodystrophy and pseudohypoparathyroidism type 1A or 1B [23]. Behavioral abnormalities have been observed in human imprinting disorders and in various mouse models in which imprinted genes have been mutated. For example, the obesity of Prader-Willi-syndrome patients is, at least in parts, a result of an impaired eating behavior. Knock-out studies in mouse showed that the two paternally expressed *Peg1* and *Peg3* genes have a clear behavioral phenotype [24]. Females that inherit a null allele for these genes from their fathers behaved 'deficiently' with respect to maternal care behavior including placentophagy and nest-building.

In this thesis, we will discuss the imprinting phenomena in details and investigate the association of imprinted genes with cell differentiation processes, namely hematopoiesis or blood cell development.

1.5.1.1.2 Histone modifications

Histones are proteins that act as a spool around which DNA can wind. When specific amino acids of histones are modified with chemical tags (acetylation, methylation, and phosphorylation), these tags can influence the physical shape of chromatin structure, which in turn, determines the accessibility of the associated chromosomal segment for binding to DNA-binding proteins (transcription factors) [11]. If chromatin is condensed (heterochromatin structure), DNA transcription doesn't occur and related genes will be inactive. If chromatin is relaxed, DNA will be easily accessible and can be transcribed and being active (Figure 1-6).

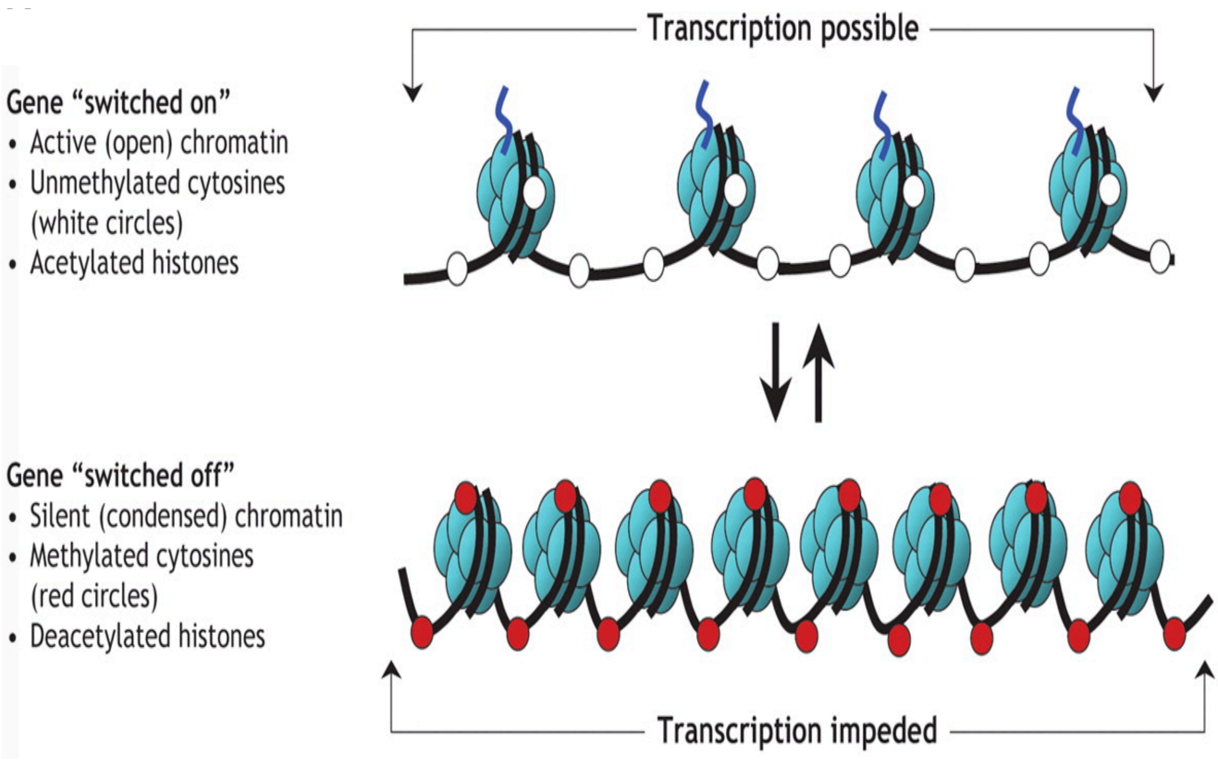

Transcription possible
Gene "switched on"
• Active (open) chromatin
• Unmethylated cytosines (white circles)
• Acetylated histones
Gene "switched off"
• Silent (condensed) chromatin
• Methylated cytosines (red circles)
• Deacetylated histones
Transcription impeded

Figure 1-6 Schematic of the reversible changes in chromatin organization that influence gene expression. Genes are expressed (switched on) when the chromatin is open (active), and they are inactivated (switched off) when the chromatin is condensed (silent). White circles = unmethylated cytosines; red circles = methylated cytosines. Source: modified from [25].

1.5.2 Post-transcriptional regulation

Post-transcriptional regulation is the control of gene expression at the RNA level, i.e after the transcription of a gene into mRNA and before the translation of RNA into a protein. The cellular events occurring at that level of regulation rely on specific RNA–protein interactions that either result in the targeted degradation of the mRNA or inhibit the translation process to make proteins [26]. Gene expression can be controlled at this level through the following mechanisms:

1.5.2.1 mRNA processing, stability, and degradation

Gene expression can be controlled by changes in pre-mRNA processing and alternative splicing, which produces various mRNA forms by removing different combinations of introns based on which proteins are needed by the cell. Also, changes in mRNA stabilities contribute to the overall regulation of gene expression. Some mRNAs in eukaryotic cells are stable and have half-lives of more than 10 hours. Many, however, have half-lives of few minutes or less. These unstable mRNAs often code for regulatory proteins, such as growth factors and transcription factors, whose production rates need to change quickly in cells [27]. Cheadle et al.2005 investigated the effect of changes in mRNA stability on gene expression during T cell activation using microarray experiments. They concluded that regulation of mRNA stability contributes significantly to the observed changes in gene expression in response to external stimuli [28]. In the same context, by binding to certain regulatory molecules like RNA binding proteins (RBP), mRNA will be directly or indirectly degraded or sequestrated in P-bodies for storage.

1.5.2.2 Interaction with non-coding RNAs

Noncoding RNAs such as microRNA (miRNAs) and long non-coding RNA (lncRNAs) have gained extensive attention in recent years as a potentially new and crucial layer of post-transcriptional biological regulation [29].

miRNAs are small non-coding RNA molecules of about 22 nucleotides that have been characterized in virtually all animals and plants. miRNAs are transcribed from different genomic loci, which implies their regulation by other transcription factors [30]. These genomic loci encode for long RNAs with a hairpin structure that when processed (cleavage) by a series of enzymes (Drosha and dicer) synthesizes a miRNA duplex of 22 nucleotides [31]. miRNAs often repress target genes through translational silencing of the mRNA or through degradation of the mRNA, via complementary binding to specific sequences in the 3' UTR region of the target gene's transcript [32] (Figure 1-7).

A miRNA can target a plethora of mRNAs, creating a post-transcriptional regulatory network [33] that has a critical role not only in cellular functions [34] but also in pathological processes [35] especially in human cancerogenesis [33, 36-38]. A considerable amount of literature has been published on miRNA-related mutations and on the impact of somatic mutations on miRNA functions. These studies have reported that genetic variants within miRNAs or their target sites can alter miRNA function in cancers [39-43] and have been associated with cancer risk, treatment efficacy and patient prognosis [39], as well as genomic phenotypes [44].

With respect to cell differentiation, miRNAs are substantial components of the molecular circuitry that controls blood cell differentiation and determines hematopoietic lineage commitment [45].

Long non-coding RNAs (lncRNA) are non-protein coding transcripts longer than 200 nucleotides [29]. Similar to the regulatory role of miRNAs, lncRNAs control various aspects of mRNA at the post-transcriptional level. lncRNAs have a repressing regulatory effect when they bind to mRNA and facade key elements with mRNA required for processing, splicing, and translation. In other inducing or activation scenarios, lncRNAs can absorb and bind to the miRNA molecules enabling mRNA to be translated [46].

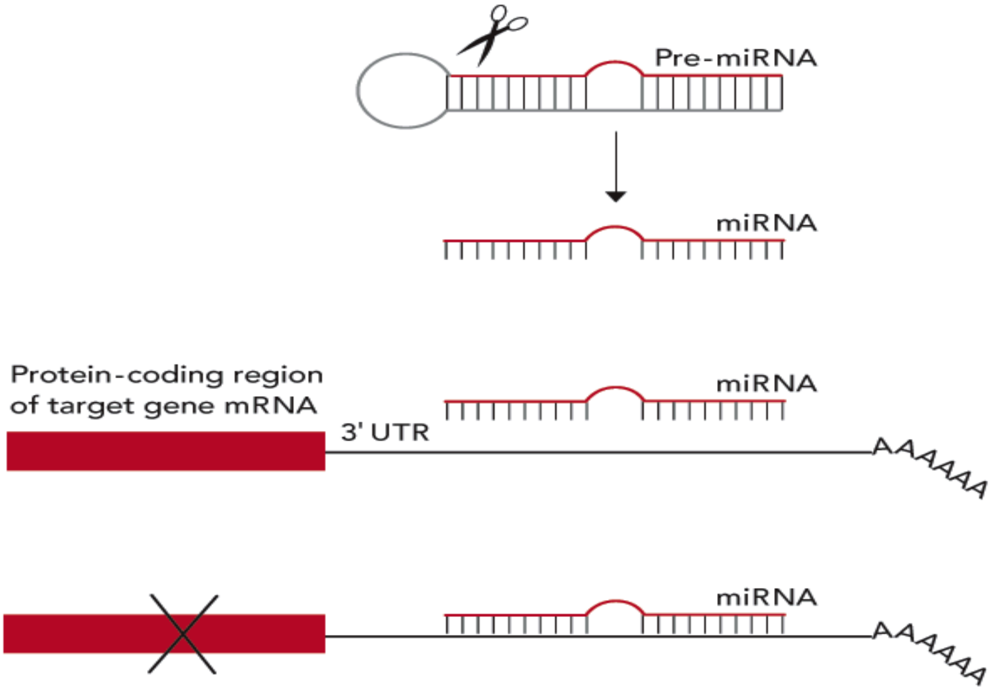


Figure 1-7 post-transcriptional regulation by miRNA interactions.
The illustration shows how a microRNA (miRNA) silences genes. It is cut out of a precursor hairpin-shaped pre-miRNA to form a mature miRNA, which binds to the 3' untranslated region (3' UTR) of a target gene's messenger RNA and turns off its activity. Source: http://www.laskerfoundation.org/awards/2008_b_description.htm.

Other non-coding RNAs such as piwi-interacting RNAs (piRNAs), endogenous siRNAs, and intron-derived miRNAs (miRtrons), were recently discovered and, yet, their regulatory roles were not deciphered. This will open new avenues of research in the field of RNA biology and, hence, will have a significant role in better understanding human development and complex diseases.

1.5.3 Post-translational regulation

Post-translational regulation refers to the control of the levels of active proteins during and after protein biosynthesis and therefore limiting their functions and stability [47]. This is achieved using two mechanisms:

1.5.3.1 Chemical modifications

Amino acid side chains may be chemically modified by attachment of chemical groups such as phosphate, acetate, amide, or methyl. Their addition or deletion may have severe effects on protein structure and function. The presence or absence of such chemicals can put also proteins in inactive state.

1.5.3.2 Degradation

Degradation refers to the life span of a certain protein. Some proteins are used in cells only for short times of e.g. a few minutes while others can last much longer. This is often controlled by protein tags like ubiquitin, which is recognized by degradation mechanisms.

These post-translational regulatory events are essential mechanisms used by eukaryotic cells to diversify their protein functions. Imperfections in these post-translational events can lead to numerous developmental disorders and human diseases [48]. Recently, Wang et al. 2014 revealed the critical role of the post-translational events (glycosylation, phosphorylation, acetylation and methylation) in the regulation of the pluripotency state of human cells [48].

1.5.4 Other factors affecting the regulation machinery

In addition to the aforementioned cellular events and genetic factors affecting the regulatory machinery, there are other extrinsic factors which take part in regulating specific biological processes. For instance, hematopoiesis (the blood cell differentiation process) is regulated in part by extrinsic signaling molecules including colony-stimulating factors (CSFs) and interleukins (ILs) that activate intracellular signaling molecules such as kinases and cytokines. These subsets of factors are known to influence Hematopoietic Stem Cell (HSC) pluripotency, proliferation, and lineage commitment (www.RnDSystems.com/HSC).

Moreover, it has been shown that biophysical properties including wettability, surface topography, and surface chemistry, could also affect the biological performance of human embryonic stem and induced pluripotent stem cells [49].

1.6 Motivation and goal of the work

Given the highly complex functional interdependencies between the molecular components (such as genes, TFs, and miRNAs) and mutations thereof in a living cell, biological processes as well as disease pathogenesis are rarely a consequence of the activity of a single molecule, but typically reflect a combination of interactions between the associated regulators (TFs and miRNAs) and their target genes [8, 50]. Such cellular interactions occurring on multiple genomic levels compose a complex and densely connected regulatory machinery. Uncovering the architecture, dynamics, and the interwoven nature of the regulatory machinery on different levels of regulations remains a challenging task and a focal point of modern systems biology.

Moreover, the correct identification of the combined regulatory interactions on different levels of regulation, will not only help in labeling hundreds of genetic molecules that are responsible for diseases, but also in identifying disease-associated cooperative functional modules of different genetic molecules. This would improve our understanding of disease development, diagnosis, and in turn, would suggest novel therapeutic strategies in disease treatment. However, this depends largely on integrating information from biological knowledge bases and large-scale omics data from different sources and experiments that capture the regulatory events occurring on the levels of regulations.

To this end, we aim in this work at developing practical computational approaches that integrate heterogeneous genomic datasets to unravel the combined regulatory interactions between different molecular elements. Then, we applied these approaches to omics data for human breast cancer as well as for hematopoiesis as a well-established model for stem cell differentiation.

1.7 Author contributions

The contribution of this thesis is two-fold. First, we present integrative systems biology approaches and bioinformatics frameworks that we have developed for reverse engineering the gene regulatory networks at transcriptional and post-transcriptional levels from heterogeneous genomic data sources. Second, we apply our approaches to hematopoietic datasets as an example for cellular differentiation and to breast cancer as well as diabetes datasets as examples for complex diseases. Then, we discuss our results and the potential biological findings and conclusions.

The work presented in this dissertation has led to the following publications and conference posters. At the beginning of each chapter, we list the corresponding publication(s) on which it is based. Furthermore, the contributions of the co-authors are reported in the text as well.

1.7.1 Publications

- Mohamed Hamed, Christian Spaniol, Maryam Nazarieh, and Volkhard Helms, TFmiR: A web server for constructing and analyzing disease specific transcription factor and miRNA co-regulatory networks. Nucleic Acids Research, 2015, doi:10.1093/nar/gkv418

- Mohamed Hamed, Christian Spaniol, Alexander Zapp, and Volkhard Helms, Integrative network based approach identifies key genetic elements in breast invasive carcinoma. BMC Genomics, 2015. 16 (Suppl 5): p. S2.

- Mohamed Hamed, Siba Ismael, Martina Paulsen, and Volkhard Helms, Cellular functions of genetically imprinted genes in human and mouse as annotated in the Gene Ontology. PLoS One, 2012. 7(11): p. e50285.

- Mohamed Hamed, Daniel Patrick Nitsche, Ulla Ruffing, Matthias Steglich, Janina Dordel, Duy Nguyen, Jan-Hendrik Brink, Gursharan Singh, Mathias Hermann, Ulrich Nubel, Volkhard Helms, and Lutz von Muller, Whole Genome Phylotyping and Microarray Profiling of nasal and blood stream Methicillin-Resistant Staphylococcus aureus isolates: Clues to phylogeny and invasiveness. Infection, Genetics and Evolution, 2015.

- Mohamed Hamed, Jonathan Odul, Andrew Steven Miller, Koichi Kawakami, and Kitamoto Asanobu. "NAS: Neuron Analyzer Suite for Automatic Analysis of Neuronal

Activities from Calcium Imaging Data. International Journal of Pharma Medicine and Biological Sciences (IJPmbs) Vol. 4, No. 3, July 2015.

- Christian Spaniol, Mohamed Hamed, Johannes Trumm and Volkhard Helms, Mebitoo: an Extensible Software Framework for Bioinformatics Analysis Workflow Automatization [BICoB-2015, 7th international conference on bioinformatics and computational biology].

- Alexander Zapp, Volkhard Helms, and Mohamed Hamed, SnvDMiR: Associating the genomic proximity of genetic variants with deregulated miRNAs and differentially methylated regions. [ICABEE -2015, 2nd international conference on Advances in Bio-Informatics and Environmental Engineering].

- Irhimeh M.R, Barthelmes D, Mohamed Hamed, Zhu L, Helms V, Gillies M.C, Shen W, Novel Gene Regulatory Network in diabetic bone marrow-derived endothelial progenitor cells. [In revision].

- Other four manuscripts are either submitted or in preparation.

1.8 Organization of the thesis

This dissertation is organized in nine chapters, including this introduction chapter (Chapter 1) and the conclusion (Chapter 9). At the beginning of each chapter, we list the related scientific articles and a synopsis, which provides a general overview about the chapter contents in the context of the thesis.

In chapter 2, we will discuss the state–of-the-art of existing methods that touch the same research problem and we will briefly illustrate the used genomic data repositories, the applied statistical methods, and the involved biological resources in this work.

Chapter 3 presents the main body of software development achieved in this research work. It focuses on the development of scalable computational approaches and automated systems biology tools and pipelines to integrate experimentally acquired and computationally derived omics data and unravel the combinatorial regulatory interactions between different molecular elements.

First, we present an automated GRN pipeline that constructs the genomic regulatory machinery of a cell from expression, sequencing, and annotation datasets through three modules implemented as separated software components (plugins) that are hosted by our software framework Mebitoo for automation of bioinformatics workflows. Then, we further extended it to a general integrative network-based approach that involves also post-transcriptional interactions and reports the computational analysis of gene and miRNA transcriptomes, DNA methylome, and somatic mutations. This aimed at identifying putative disease drivers and novel targets for therapeutic treatment. With regard to the incorporation of somatic mutations with other genomic data sets, a stand-alone pipeline named "SnvDMiR" was implemented to explore possible genomic proximity relationships between somatic variants and both differentially methylated CpG sites as well as differentially expressed miRNAs. With respect to genomic mutations, we also present an NGS pipeline and apply it to two groups of bacterial isolates (nasal and invasive) to investigate the phylogenetic positions of the recently emerged t504 clone (Spa-type t504) in the Saarland province of Germany and to better understand the infectivity mechanism of the invasive group. Motivated by all of this, we developed TFmiR as a freely available web server for deep and integrative downstream analysis of combinatorial regulatory interactions between TFs/genes and miRNAs that are involved in human disease pathogenesis.

In the second part of this thesis (from chapter 4 to chapter 8), we will demonstrate the usefulness and applicability of the developed approaches and frameworks by applying them to two different cellular activities: hematopoietic stem cell differentiation and disease pathways.

In chapter 4, we discuss genomic imprinting as an epigenetic phenomenon that is closely associated with cell development and cellular differentiation. We characterize the role of imprinted genes during differentiation processes and comprehensively investigate the cellular functions of the whole set of imprinted genes, paternally expressed genes, and maternally expressed genes as well as the transcription factors that are predicted to regulate the imprinted genes and their relatedness to cell differentiation in both human and mouse. The findings of this chapter motivated the study presented in chapter 5 regarding the nature and extent of the role of imprinted genes in hematopoietic stem cell differentiation.

In chapter 6, we demonstrate the effectiveness of one of our developed approaches (the integrative network-based approach) to identify genetic key elements that could possibly drive the tumorigenesis in human breast cancer. Also in chapter 7, we will consider the differential network analysis concept that makes use of our developed GRN pipeline to elucidate the molecular mechanisms by which diabetes impairs Bone marrow-derived endothelia progenitor cells (EPC) in mouse.

Chapter 8 concerns the role of genomic mutations. Here, we apply our implemented NGS pipeline to identify core-genome SNPs and genetic variations between two phenotypic groups in a similar analogy to somatic mutations between the healthy and disease cohorts. This project involved two groups of MRSA bacterial isolates (nasal and invasive) to investigate the phylogenetic positions of the recently emerged t504 clone (Spa-type t504) in the Saarland province of Germany and to better understand the infectivity mechanism of the invasive group as a prototype example for “from genotype to phenotype” studies. In the last chapter (chapter 9), we summarize the results achieved in this thesis and discuss the current limitations of the introduced approaches and directions for further improvements and outlook. Finally the appendices A, B, C, D, and E contain supplementary information for chapters 4, 5, 6, 7, and 8, respectively.

In summary, the work presented in this thesis has led to the development of interesting computational approaches that are introduced to the scientific community as non-commercial software toolkits. The provided topological and functional analyses of our approaches as validated on cellular differentiation and complex diseases promote them as reliable systems biology tools for researchers across different life science disciplines.

2. Theory and computational biology tools

Synopsis

In this chapter, we discuss the state–of-the-art of existing methods that touch the same research problem. A brief description of some of the data repositories, statistical tools, and biological resources used in this work is also provided here.

2.1 GRN reconstructing methods

A gene regulatory network is typically represented as a graph in which the nodes are target genes or regulators (TFs, miRNAs) and the edges between nodes represent gene interactions through which the products of one gene affect those of another. These regulatory links can be inducting or activating, where an increase in the expression of one leads to an increase in the other, or inhibitory/ repressing, where an increase in one leading to a decrease in the other. A series of edges indicates a chain of such dependencies, with cycles corresponding to feedback loops.

By reconstructing the gene regulatory network (GRN) of a single cell or of a multicellular system we mean here the process of unraveling the regulatory machinery of this biological system and then studying its structure, function, and mode of operation. Over the past years, several methods have been developed and applied to reconstruct GRN topologies from high- throughput data sources. In the next section, we provide an overview of these methods that are categorized according to the underlying model of gene regulation.

2.1.1 Boolean Networks

Boolean networks are one of the oldest dynamical methods that generate experimental time series for gene expression of gene circuits[51]. The state of each variable (genes) at the next time step depends in a deterministic manner on the states of some other variables at the current time step. These dependencies are encoded in the form of matrix-like condition tables. Boolean networks are based on the assumption that binary on/off switches in discrete time steps can describe important aspects of gene regulation. In a boolean network, the network state can be defined as *n*-tuples of 0s and 1s describing which genes in the network are or are not expressed at a particular moment. For a network of *n* genes, there are 2^*n* possible different states. As time progresses, the network states transition through this 'state space', switching from one state to another. These states can be monitored to determine which states have been reached and which (cycles between) states the network prefers to stay in once they are reached (so-called attractors). Stochastic and probabilistic extensions to Boolean networks were also proposed by Akutsu et al. [52] and Shmulevich et al. [53], respectively.

2.1.2 Dynamical models

Dynamical models such as ordinary differential equations (ODE) are important classes of GRN inference methods and probably the most-used formalism for modeling genetic networks. In these models, the concentrations of genes, mRNAs, or proteins are represented by continuous, time-dependent variables as follows:

$$\frac{dx_i}{dt} = f(\mathbf{x})$$

where x_i is the expression level of gene i and x is the state vector containing the expression levels of all other genes. The so-called input functions f can be linear (first order) or non-linear ODE functions. The linear model is the most commonly used dynamical model for gene network inference due to its simplicity [54-56]. However, it turned out that linear models often don't provide plausible results when only mRNA concentrations are modeled. Successful models often also requires considering the process of protein translation via introducing protein concentrations as further dynamical variables. On the other hand, methods of reconstructions using non-linear models employ complicated numerical optimization techniques to fit experimental gene expression data [57, 58]. Yip et al. presented in 2010 ODE models for knockout and perturbation data sets to infer the topology of GRN networks and achieved the best score in the Dream 3 challenge [59].

2.1.3 Stochastic approaches

Although differential equations allow predicting the exact concentrations of genes and proteins, they assume that these molecular concentrations vary continuously and deterministically. However, in real biological systems, cellular activities and regulatory processes are stochastic processes subject to considerable noise [60, 61]. Especially when the number of molecules in a certain cellular reaction is small, stochastic methods can be efficient in modeling the underlying networks [62]. Due to the complexity involved in estimating, solving and analyzing stochastic models, these are rarely used to model real networks of more than two or three genes and therefore are not applicable to high-throughput datasets of the sort considered in this thesis.

2.1.4 Bayesian Networks

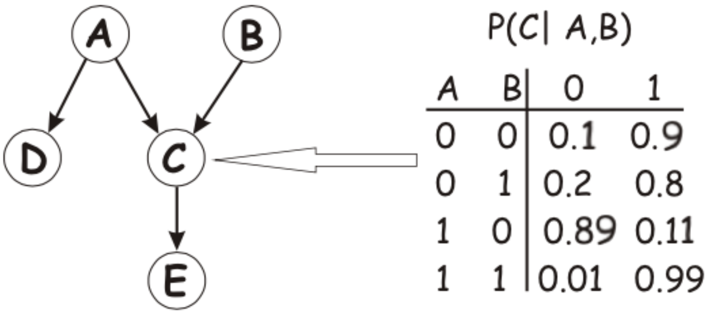
Among the various kinds of computational methods that have been presented for reconstructing gene networks, Bayesian network (BN) approaches have shown great promise to infer causal relationships between genes based on their expression profiles. BNs are in principal probabilistic graphical models that encode dependencies between genes and represent the state of each gene as a joint probability distribution via a product of terms [63].

Theoretically, Bayesian networks are graphical notation for conditional independence assertions between random variables and hence for compact representation of full joint distributions. Consider a finite set $X = (X_1, X_2.... X_n)$ of random variables (genes) and their values refer to their gene expression measurements. The joint distributions over the set X can be represented as a product of conditional probabilities, where each variable X_i is associated with a conditional probability $P(X_i \mid Pa_i)$ and $Pa_i \subseteq X$ is a set of variables that are the parents of X_i. Thus, the values the X_i are independent on the values of the other variables given the parents of $X_{i,}$ in other words, the parents of X_i directly influence the values for X_i (an example is shown in Figure 2-1).

$$\Pr(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i \mid \mathbf{Pa}_i).$$

The theory of learning Bayesian network structure from data can be formulated as follows:

Given a training set $D = X^1, X^2 \dots X^N$ of independent instances of X, find a network $B = (G, \Theta)$ that best matches D, where G is the network structure and Θ describes the graph parameters of the conditional probability table for each random variable X.



P(C| A,B)

A	B	0	1
0	0	0.1	0.9
0	1	0.2	0.8
1	0	0.89	0.11
1	1	0.01	0.99

Figure 2-1 Bayesian network representation.
Left: acyclic directed graph showing a Bayesian network with five random variables, where nodes (*A* to *E*) represent genes and edges represent direct dependencies between them. Right: conditional probability distribution for the gene *C*, where the expression level of its parents is discretized to a Boolean value. The product form specified by this Bayesian network is $P(A, B, C, D, E) = P(A)\,P(B)\,P(C \mid A, B)\,P(D \mid A)\,P(E \mid C)$.

The common method of Bayesian structure learning is to start from an initial network model (usually known as prior knowledge) and then adding nodes/genes using some operators (add edge, reverse edge, and remove edge). At every iteration, cyclic structures are removed and only candidate structures or models are subjected to the next step. Then, a structure-scoring function is applied to find which model best fits the data. Since the number of generated models is super-exponential in the number of involved random variables (genes), an efficient search algorithm (e.g. greedy algorithm) has to be employed to search the model space and find the model with the highest score. The approach can be summarized in the following procedure:

```
I/P
        -M measurements of a finite set of random variables $D = X_1, X_2 \dots X_n$.
        -An initial network $N_{init}(V, E)$ where $V \in D$.

O/P
        -A directed Bayesian network $B = (G, \Theta)$ that best matches $D$, where $G$ is the network
         structure and $\Theta$ describes the parameters of the conditional probability table for each $X_i$.

Procedure:

For each variable $X_i \in D$ and $\notin V$:

        Step 1:  Add $X_i$ to $N_{init}$ using the operators (add edge, reverse edge, and remove edge) and
                 generate a structure space $S = \{S_1, S_2, \dots\dots S_h\}$ that contains all possible candidate
                 structures.
        Step 2:  Remove cyclic structures.
        Step 3:  Apply a scoring function (ex: $BDe$) to all candidate structures.
        Step 4:  Search for the highest score structure $S_{best}$ using an effective search algorithm.
        Step 5:  Update the initial network with the inferred best structure, $N_{init} = S_{best}$.
```

A commonly used scoring function for evaluating the candidate Bayesian models is called Bayesian Dirichlet method (BDe) scoring metric which calculates the posterior probability of a network G given data D [64]. The posterior probability of a graph given the data is:

$$BDe(G : D) = log\ P(G|D) = log\ P(D|G) + log\ P(G) + C$$

where $P(D|G)$ is the marginal likelihood which averages the probability of the data D over all possible parameter assignments to G, $P(G)$ is the prior probability of network G, and C is a constant independent of G [64].

An advantage of using Bayesian networks to model gene regulatory networks is that they can readily handle the stochastic aspects of input data as well as noisy and incomplete datasets. These problems are typical for gene expression data. Moreover, Bayesian networks facilitate the combination of prior or domain knowledge and input data. This allows making use of what is already known from regulatory interactions and regulatory repositories to infer regulatory relationships between genes involved in specific input data. Also, Bayesian networks encode the strength of causal relationships with probabilities. Therefore, prior knowledge and data can be combined with well-studied techniques from Bayesian statistics [65]. In addition, Bayesian methods offer an efficient approach for avoiding the overfitting of data. There is no need to dedicate some of the available data for testing. Using the Bayesian approach, models can be evaluated or scored in such a way that all available data can be used for learning. On the other hand, Bayesian networks don't model the dynamical aspects of gene expression. Since the Bayesian networks are directed acyclic graphs (DAGs), this doesn't allow network structures such as feedback loops or self–regulation links to be modeled, which is actually the case for most human genes (that typically have a negative auto-regulatory feedback resulting in a sigmoidal thresholding of their maximal expression levels).

2.2 Biological data repositories

2.2.1 The Cancer Genome Atlas (TCGA)

The TCGA portal [66] (http://cancergenome.nih.gov/) is a cancer-specific data warehouse to search, download, and analyze consistent genome-scale datasets generated from cancer patient samples by the TCGA consortium. The TCGA initiative was established in 2005 in the context of the "war on cancer" initiative. Most data samples are freely available to allow researchers around the world to analyze and make predictions. In this thesis, TCGA healthy and tumor samples were analyzed using different techniques such as gene expression profiling, genome wide DNA methylation, miRNA profiling, and SNP genotyping. Every TCGA sample carries a unique barcode, which is composed of a set of data element identifiers such as tissue source site, patient name, and sample type (healthy or tumor). See Figure 2-2. In chapter 6, we applied our computational pipeline to breast cancer data downloaded from the TCGA portal and revealed strong associations between regulatory elements from different genomic data sources.

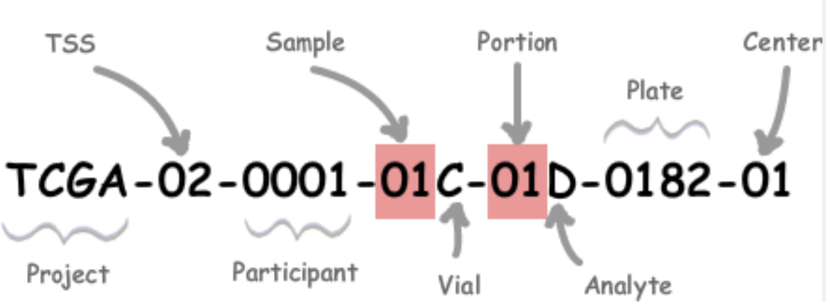


Figure 2-2 Illustration of the TCGA barcode and its data element identifier.
Source: National cancer institute (NCI), https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode.

2.2.2 Gene Expression Omnibus (GEO)

GEO (http://www.ncbi.nlm.nih.gov/geo/) is an international public genomic repository maintained by NCBI to collect and freely disseminate raw and preprocessed microarray and next generation sequencing datasets. GEO data are organized in three entities:

1- Platform: a platform is a list of probes related to an array technology provider.
2- Sample: a sample describes the set of molecules (here: genes) whose expression profiles are measured in certain condition/tissue.
3- Series: a series groups samples into meaningful data sets, which make up an experiment.

The hematopoiesis study presented in chapter 5 used expression profiles of blood cell lines downloaded from GEO.

2.3 Biological knowledge databases

Biological databases are informative digital libraries collected from scientific experiments, published literature, and computational analyses of high-throughput data. These biological data are often structured and represented in tabular data, XML formats, key-delimited records, ontology classes, well-established attributes, and relationships. Various biological databases were tightly integrated in the tools and approaches developed in this thesis. Below, we will briefly introduce some of them below.

2.3.1 Gene Ontology (GO)

The Gene ontology [67] is a set of formal vocabularies and explicit specifications of gene annotation terms that are used to describe the attributes of genes in an organism. GO is composed of three sub ontologies on the biological processes (BP), molecular functions (MF), and cellular components (CC) annotated to genes. The building blocks of GO are the terms (also called functional classes or functional categories). Each GO term has a unique ID and a textual name, Ex, GO: 0042660: `positive regulation of cell fate specification`. A gene can be associated with one or more GO terms and may belong to different GO sub ontologies. The terms of the GO database are organized in a hierarchical structure where a few general terms such as *developmental process* are linked to numerous more specific terms on the next hierarchical level. Note that cycles are allowed. Recently, the developers team behind the David resource [68] has established GO_FAT. This is a subset of the full set of GO terms so that the broadest terms should not overshadow more specific terms.

2.3.2 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a group of linked comprehensive databases that store information about gene products, biological pathways, drugs, diseases, and chemical substances as a knowledge base for systematic analysis of genetic molecules [69]. The core of the KEGG suite is the KEGG PATHWAY database that contains metabolic pathway maps integrating genes, proteins, miRNAs, and chemical compounds as well as disease genes and drug targets.

Similar to the Gene Ontology, KEGG can be used to determine whether or not a set of genes is functionally enriched. However, the term (enrichment) here is more related to the biological pathways, in other words, the contributions of these genes in the background of all chemical reactions occurring within the cell such as metabolism, membrane transport, and signal transduction.

2.3.3 Molecular signature database (MSigDB)

MSigDB is a well-annotated gene set representing the universe of biological processes and is used for interpretation of large-scale genomic data [70]. Genes are grouped into annotated sets based on specific genomic properties among them such as shared binding sites for transcription factor families (motif gene sets), associations with the same GO functional terms (GO gene sets), or being involved in the same diseases (oncogenic signature sets). Here, we incorporated the motif gene sets into our developed tools and used them also to investigate whether binding sites for distinct TFs are enriched in the promoter regions of imprinted genes (chapter 4).

2.3.4 Regulatory interaction databases

Several databases and online repositories have been developed in order to facilitate the research on predicted and experimentally verified genome-wide transcriptional and post-transcriptional regulatory interactions. For instance, TransFac [71] and MsigDB [70] maintain interactions of TFs regulating genes (TF→gene). TransmiR [30] provides information on which TFs regulate miRNAs (TF→miRNA). mirTarBase[72], TarBase [73] and miRecords [74] comprise miRNAs and their target genes (miRNA→genes) in different organisms. Although still little is known about miRNA-mediated miRNA regulations, miRNA→miRNA interactions were computationally predicted and maintained in the PmmR database [75]. An extensive study of the integration mechanisms of such databases and further downstream analysis of the involved genetic molecules and their pathways in cancer is described in Chapter 3.

2.4 Statistical tools

2.4.1 Over representation analysis (ORA)

The demand on computational biology for evaluating results of high-throughput data analysis has led to the development of several popular tools and statistical approaches [76]. ORA is a widely used statistical approach that increases the likelihood for researchers to identify biological terms most relevant to their study. More specifically, ORA compares a reference set of genes to a test set in terms of their associations with a certain biological term. For instance, when considering a certain GO functional term, this method assesses whether this GO term is over-represented or under-represented in the respective study set and estimates how likely this is to happen by chance. DAVID [68] and WebGestalt [77] are famous examples for tools that have been developed for the purpose of ORA analysis. ORA uses both parametric statistical tests (ex: hypergeometric test) and non-parametric statistical tests (ex: Kolmogorov-Smirnov test) to assess the significance of term enrichment. ORA was a central part in the study of chapter 4 where we examined the cellular functions and motif enrichment of imprinted genes and hence their regulatory roles in cellular differentiation and blood cell development (chapter 5).

2.4.2 Hypergeometric test

The hypergeometric-based test estimates the discrete probability that describes the number of successes in a sequence of *n* draws from a finite population without replacement.

Given a set of *N* study genes, of which *x* belong to a functional category *C*, and a population set of size *M*, of which *k* belong to *C*. Then, the probability of observing *x* genes out of *N* by chance belonging to the *C* category containing *k* genes from a total population of size *M* can be modeled as follows:

$$P = 1 - \sum_{i=0}^{x} \frac{\binom{k}{i}\binom{M-k}{N-i}}{\binom{M}{N}}$$

The closer the probability (*p-value*) is to 0, the more unlikely is the chance of error that the majority of genes belong to that category *C* (enriched).

2.4.3 Kolmogorov-Smirnov test (KS test)

KS test is a non-parametric (distribution free) test for comparing a sample to a reference probability distribution (one-sample KS test), or for comparing the distributions of two samples (two-sample K–S test). The KS test computes the distance between the empirical distribution function of a sample and the cumulative distribution of the reference hypothesis (in case of the one-sample KS test) or between the empirical distribution functions of two samples (in case of the two-sample KS test).

Due to its sensitivity to differences in both the location and shape of the empirical distribution functions, the two-sample KS test is considered as one of the most useful and most general nonparametric methods for comparing two samples. Therefore, we incorporated the KS test in our developed TFmiR web server (Chapter 3) to compare two distributions of gene pairwise functional similarity scores.

2.4.4 Multiple test correction

Since we are testing the enrichment of a set of genes with multiple biological categories, several independent statistical tests are performed simultaneously. This leads to increasing the probability for false positive predictions or what is a so-called “Type 1 error”. The multiple testing corrections adjust p-values derived from multiple hypothesis testing to correct for the occurrence of false positives. In gene expression analysis, for example, false positives could be those genes that are found to be statistically differentially expressed between two conditions, but are not in reality.

In our developed computational tools, we utilized the Benjamini and Hochberg (BH) [78] False Discovery Rate (FDR) as a multiple testing correction approach. Instead of controlling the probability of committing any type I error by setting a more severe cut off level as in the Bonferroni method, this BH method controls the expected proportion of errors among the rejected null hypotheses.

$$FDR = E\left(\frac{\text{number of falsely rejected null hypotheses}}{\text{number of rejected null hypotheses}}\right)$$

Therefore, it is the least stringent of all corrections and keeps a good balance between the discovery of statistically significant genes and limitations of the predictive power due to the occurrence of false positives.

3. Approaches and Methods

This chapter is based on the following publications:

- Mohamed Hamed, Christian Spaniol, Maryam Nazarieh, and Volkhard Helms, TFmiR: A web server for constructing and analyzing disease specific transcription factor and miRNA co-regulatory networks. Nucleic Acids Research, 2015, doi:10.1093/nar/gkv418

- Mohamed Hamed, Christian Spaniol, Alexander Zapp, and Volkhard Helms, Integrative network based approach identifies key genetic elements in breast invasive carcinoma. BMC Genomics, 2015. 16 (Suppl 5): p. S2.

- Mohamed Hamed, Daniel Patrick Nitsche, Ulla Ruffing, Matthias Steglich, Janina Dordel, Duy Nguyen, Jan-Hendrik Brink, Gursharan Singh, Mathias Hermann, Ulrich Nubel, Volkhard Helms, and Lutz von Muller, Whole Genome Phylotyping and Microarray Profiling of nasal and blood stream Methicillin-Resistant Staphylococcus aureus isolates: Clues to phylogeny and invasiveness. Infection, Genetics and Evolution, 2015.

- Christian Spaniol, Mohamed Hamed, Johannes Trumm and Volkhard Helms, Mebitoo: an Extensible Software Framework for Bioinformatics Analysis Workflow Automatization [BICoB-2015, 7th international conference on bioinformatics and computational biology].

- Alexander Zapp, Volkhard Helms, and Mohamed Hamed, SnvDMiR: Associating the genomic proximity of genetic variants with deregulated miRNAs and differentially methylated regions. [ICABEE -2015, 2nd international conference on Advances in Bio-Informatics and Environmental Engineering].

Synopsis

In this chapter, we present scalable computational approaches and automated pipelines that we implemented as freely available software tools to integrate heterogeneous sources of large-scale genomic data and to unravel the combinatorial regulatory interactions between different molecular elements. We started with a GRN pipeline to reverse engineer the regulatory interactions from gene expression and gene sequence data. Then, we expand it to a general integrative network based approach involving miRNA expression, DNA methylation, and genetic variants. An NGS pipeline was implemented to identify the core genome SNPs between two different phenotype groups in analogy to the identification of somatic mutations between disease and healthy samples. A standalone proximity pipeline was also implemented to study the vicinity relationships between a significant set of gene promoters, miRNAs and genetic variations. Finally, we developed TFmiR as a freely available comprehensive web server for deep and integrative analysis of regulatory information between TFs/genes and miRNAs and their interwoven critical roles in the pathology of human diseases.

In this chapter, we present the bioinformatics tools and approaches that were implemented in the course of this thesis. The aim of these approaches is unraveling the interwoven gene regulatory network between genetic molecules that are involved in cellular functions and disease pathways. We applied these approaches to omics data of human breast cancer as well as hematopoiesis as a well-established model for stem cell differentiation.

3.1 The model of gene regulation

As shown in Chapter two, several complex approaches to reconstruct cellular networks from gene expression data have been published over the last few years. Among these, only Bayesian networks infer causal relationships between genes while making use of known regulatory information that is already stored in regulatory databases. This information is referred to here as a primary or prior knowledge. Hence, we adopted the Bayesian approach in our implemented pipeline to reconstruct the GRN from expression and gene sequence data.

In this thesis, we formulate the architecture of GRNs as follows, both target genes and genes coding for transcription factors are represented as nodes in the network. Regulatory interactions are represented as directed edges from TF or miRNA to target genes. On the transcription level, if a gene is silenced while the methylation level of its promoter was high, we assume that gene silencing results from the increase in promoter methylation and is not due to TFs nor miRNA regulation [79]. In addition, we assumed direct individual binding of transcription factors to the regulatory site of a gene: in other words, regulation role of transcription factors complexes are not considered here. On the post-transcriptional level, we only considered the regulation of genes by miRNAs and ignored other degradation causes of mRNA transcripts. Finally, although post-translational regulation is an important aspect in the regulatory machinery in complex cellular systems, there was no chance to model it within the scope of this work for lack of time. Nevertheless, it is discussed in the last chapter as a high potential follow-up work to this dissertation.

3.2 GRN construction pipeline

The reconstruction pipeline consists of three steps:

1. Build a weighted co-expression network from gene expression data.

2. Query known regulatory interactions that are likely involved in the constructed co-expression network and do motif search using sequencing data of network genes.

3. Learn the network topology using Bayesian approach by utilizing information from step 2 as a prior knowledge.

The first step outputs an undirected network $G(V, E)$ with edge thickness representing the correlation strength between the expression profiles of the connected genes. In the second step, we examine the regulation directionality for each undirected edge e in the co-expression network by connecting to transcriptional regulatory databases and performing motif discovery analysis. If a directed interaction between the two examined nodes is found in the regulatory databases or confirmed by motif search, the corresponding undirected edge e will be updated accordingly to a directed edge $e_{d.}$ The resulting network of this step would contain both directed and undirected edges. Next, the directed sub-network $G_d(V_d, E_d)$ representing the constructed directed edges is extracted and used as a prior knowledge to statistically learn the remaining network structure via Bayesian approach in the third step. This last step takes two inputs: the prior network constructed from step 2 as well as the expression dataset of genes V_{ud} that are still involved in undirected edges. This outputs a directed causal probabilistic network that best fits the expression data of the involved genes.

Then, the final network topology is composed of directed interactions identified in step 2 as well as interactions confirmed by both step 1 (co-expression) and step 3 (Bayesian learning). Each of these steps is detailed below; see Figure 3-1 for an overview for the entire pipeline. The last block refers to the integrative network-based approach which was developed as an extension to the GRN pipeline to process information from epigenetic data and somatic mutations, see section 3.4. The three steps of the GRN pipeline are implemented as separate software modules (plugins) and hosted by our software framework Mebitoo for workflow automation (section 3.3). We note that coupling the third module is still in progress.

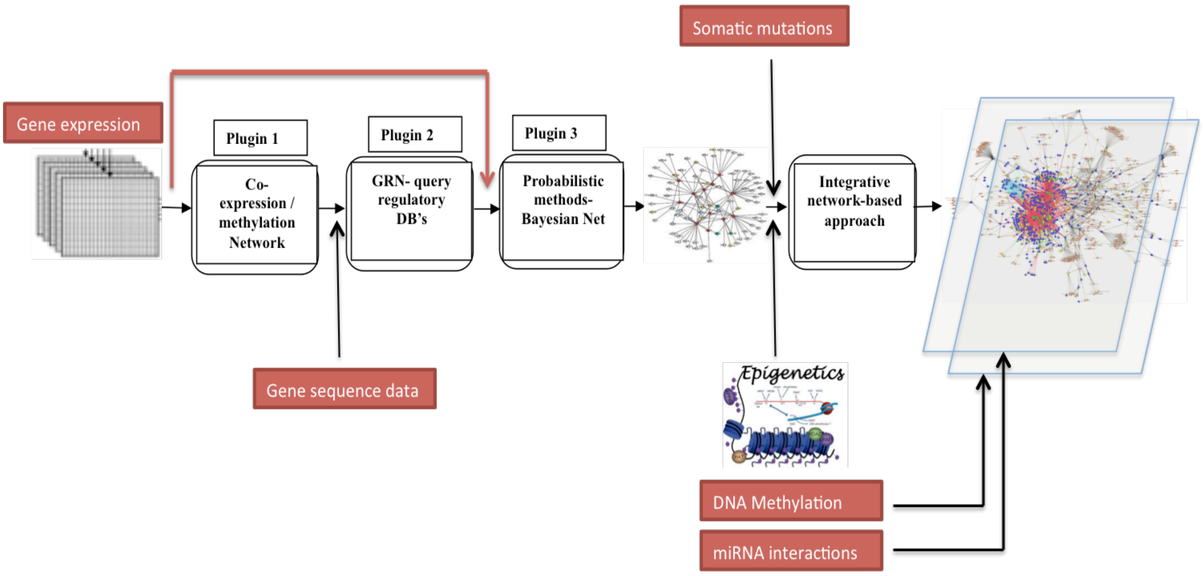
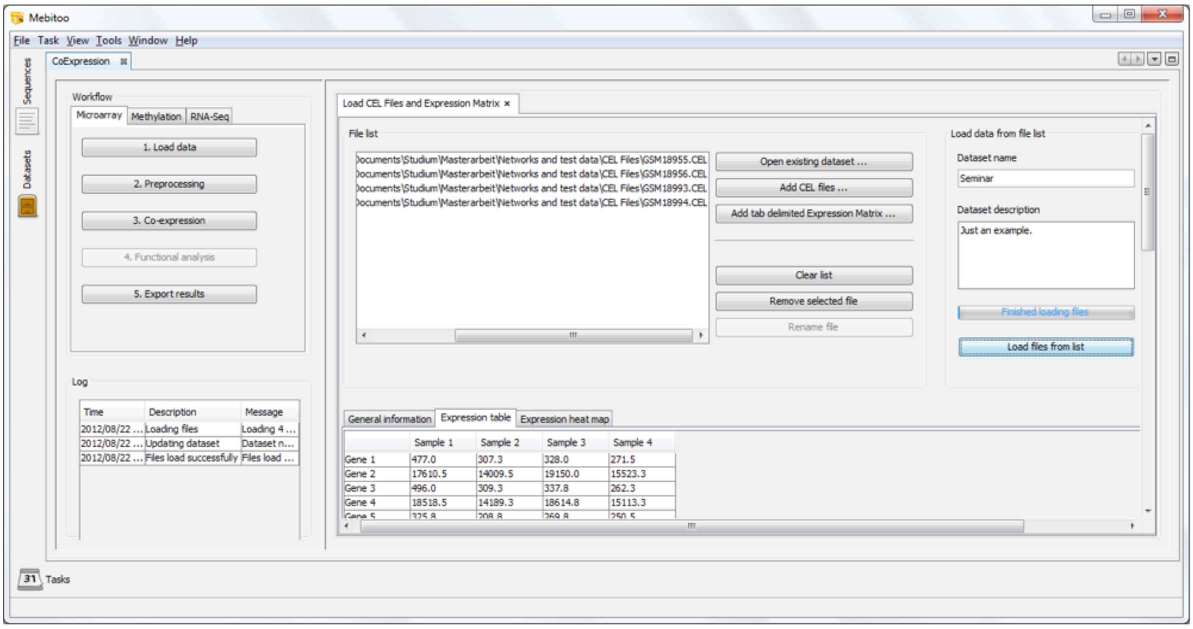

Somatic mutations
Gene expression
Plugin 1
Co-expression / methylation Network
Plugin 2
GRN- query regulatory DB's
Plugin 3
Probabilistic methods-Bayesian Net
Integrative network-based approach
Gene sequence data
Epigenetics
DNA Methylation
miRNA interactions

Figure 3-1 GRN construction pipeline from heterogeneous sources of genomics data.
The steps are as follows: (1) Build weighted co-expression network from gene expression data. (2) Query regulatory interactions and do a search for binding motifs using sequencing data of network genes. (3) Learn the network topology using Bayesian approach and utilizing information from step 2 as a prior knowledge. The last block refers to the integrative network-based approach, which was developed as an extension to the GRN pipeline to process information from epigenetic data and somatic mutations.

3.2.1 Plugin 1: Weighted co-expression network

Co-expression networks provide a widely applicable framework for assigning gene cellular functions and identifying functional network modules [80]. Gene co-expression concurs the functional similarity between genes based on gene ontology (GO) annotations [67]. Co-expression networks are defined as undirected gene networks where nodes correspond to genes, and edges between genes are determined by the pairwise similarity between gene expressions profiles and applying a particular cutoff threshold.

Users can load raw or preprocessed expression data to this plugin to start generating the co-expression network between genes (Figure 3-2). The tool offers routines for data preprocessing such as background corrections, data normalization, and probe summarization. The plugin also displays some plots such as expression heatmap, box plot, histogram plot which hint at better exploring the data before and after the pre-processing step (Figure 3-3).



File list

```
Documents\Studium\Masterarbeit\Networks and test data\CEL Files\GSM18955.CEL
Documents\Studium\Masterarbeit\Networks and test data\CEL Files\GSM18956.CEL
Documents\Studium\Masterarbeit\Networks and test data\CEL Files\GSM18993.CEL
Documents\Studium\Masterarbeit\Networks and test data\CEL Files\GSM18994.CEL
```

Load data from file list

Dataset name: Seminar

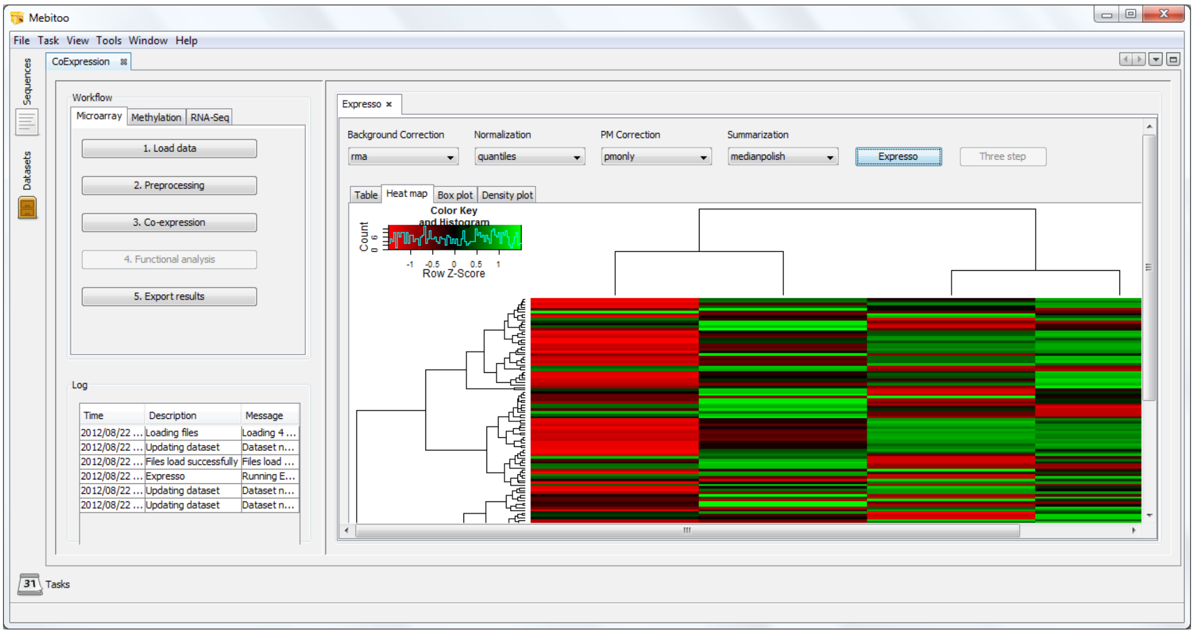
Dataset description: Just an example.

	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	477.0	307.3	328.0	271.5
Gene 2	17610.5	14009.5	19150.0	15523.3
Gene 3	496.0	309.3	337.8	262.3
Gene 4	18518.5	14189.3	18614.8	15113.3
Gene 5	325.8	208.8	269.8	250.5

Log

Time	Description	Message
2012/08/22 ...	Loading files	Loading 4 ...
2012/08/22 ...	Updating dataset	Dataset n...
2012/08/22 ...	Files load successfully	Files load ...

Figure 3-2 Data loading panel in the co-expression plugin.



Background Correction: rma

Normalization: quantiles

PM Correction: pmonly

Summarization: medianpolish

Log

Time	Description	Message
2012/08/22 ...	Loading files	Loading 4 ...
2012/08/22 ...	Updating dataset	Dataset n...
2012/08/22 ...	Files load successfully	Files load ...
2012/08/22 ...	Expresso	Running E...
2012/08/22 ...	Updating dataset	Dataset n...
2012/08/22 ...	Updating dataset	Dataset n...

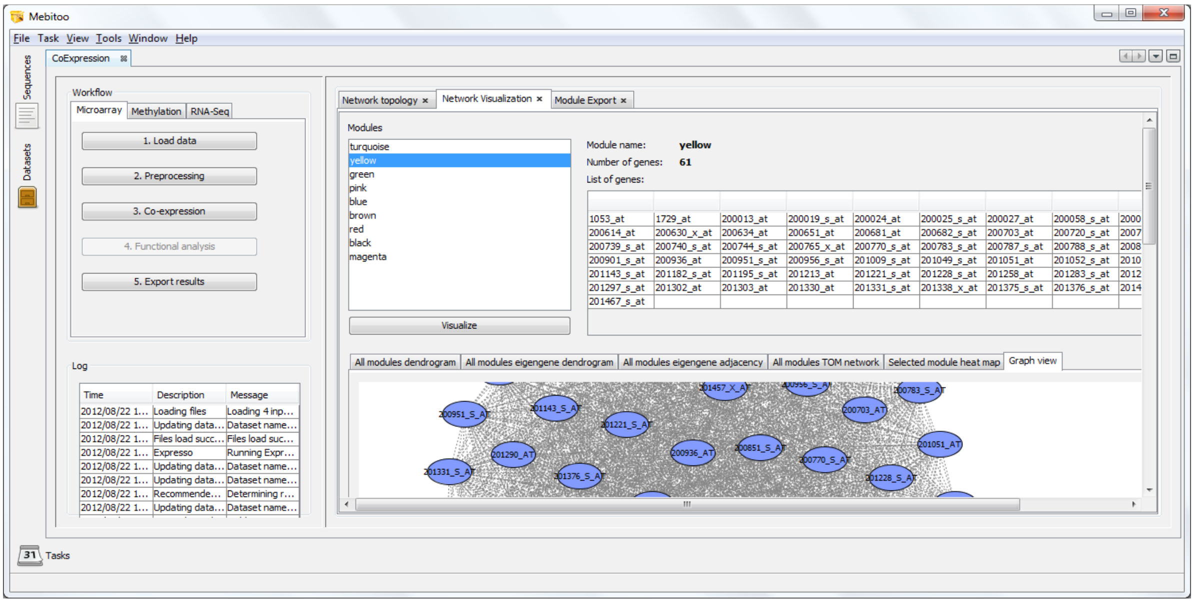
Figure 3-3 Preprocessing options of raw expression data.

The plugin utilizes the WGCNA R package [81] to build a weighted co-expression network from gene expression data. First, we measure the concordance between gene expression profiles using Pearson correlation. Then, the pairwise correlation matrix is subjected to power adjacency function to obtain a weighted correlation matrix which emphasizes high correlations at the expense of low correlations as follows.

$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

where a_{ij} is the weighted correlation that refers to the connection strength between gene pairs x_i, x_j, while β is a coefficient that controls the soft threshold curvature and its value is recommended by the tool for each dataset.

Next, we use average linking hierarchical clustering to cluster genes into co-expression network modules. Finally, for each module we display the corresponding weighted co-expression network and the list of involved genes. Results can be exported to network files and used as input parameters for the next plugin (Figure 3-4).



Module name: **yellow**

Number of genes: **61**

List of genes:

1053_at	1729_at	200013_at	200019_s_at	200024_at	200025_s_at	200027_at	200058_s_at	2000
200614_at	200630_x_at	200634_at	200651_at	200681_at	200682_s_at	200703_at	200720_s_at	2007
200739_s_at	200740_s_at	200744_s_at	200765_x_at	200770_s_at	200783_s_at	200787_s_at	200788_s_at	2008
200901_s_at	200936_at	200951_s_at	200956_s_at	201009_s_at	201049_s_at	201051_at	201052_s_at	2010
201143_s_at	201182_s_at	201195_s_at	201213_at	201221_s_at	201228_s_at	201258_at	201283_s_at	2012
201297_s_at	201302_at	201303_at	201330_at	201331_s_at	201338_x_at	201375_s_at	201376_s_at	2014
201467_s_at								

Log

Time	Description	Message
2012/08/22 1...	Loading files	Loading 4 inp...
2012/08/22 1...	Updating data...	Dataset name...
2012/08/22 1...	Files load succ...	Files load suc...
2012/08/22 1...	Expresso	Running Expr...
2012/08/22 1...	Updating data...	Dataset name...
2012/08/22 1...	Updating data...	Dataset name...
2012/08/22 1...	Recommende...	Determining r...
2012/08/22 1...	Updating data...	Dataset name...

Figure 3-4 visualizing the network and gene lists for each co-expression module.

3.2.2 Plugin 2: Online query for regulating links and motif search

This plugin was designed by the author of this thesis and implemented by Mr. Johannes Trumm during his M.Sc. thesis under the supervision of the PhD author. The co-expression network constructed in the first plugin is subjected to plugin 2 as an input parameter. This plugin matches the co-expression interactions with regulatory information retrieved from the Transcriptional Regulatory Element Database (TRED) [82], Molecular Signatures Database (MSigDB) [70], and JASPAR database [83]. Also the tool utilizes the NCBI (http://www.ncbi.nlm.nih.gov/) and HGNC (http://www.genenames.org/) repositories to download gene promoter sequences and map the input gene names to unique identifiers, respectively. Figure 3-5 shows the integrated genomic resources and software components in this plugin. Finally, the user has the option to set the parameters required for motif search and promoter region identification via a user control panel (Figure 3-6). The matching process can be summarized in the following steps:

1. Assigning transcription factors.
   All genes involved in the co-expression network and listed in at least one of the above databases to code for a transcription factor (TF) were marked as TFs.

2. Adding known regulatory interactions.
   For each TF-gene link in the co-expression network, we searched whether the databases contain a known regulation for this TF-target gene pair. In each of these cases, a directed edge was added between the transcription factor and the target gene.

3. Searching for known binding motifs.
   Here, we used the Motif Statistics and Discovery (MoSDi) [84] software to conduct a motif search for all known binding motifs of the TFs represented in the co-expression network against the promoter regions of all genes in the network. If a match was found, a new directed edge from the TF to the gene was added. Finally, the constructed directed interactions are visualized in an interactive display and can be exported to Cytoscape [85] or VisANT [86] network files (Figure 3-7).

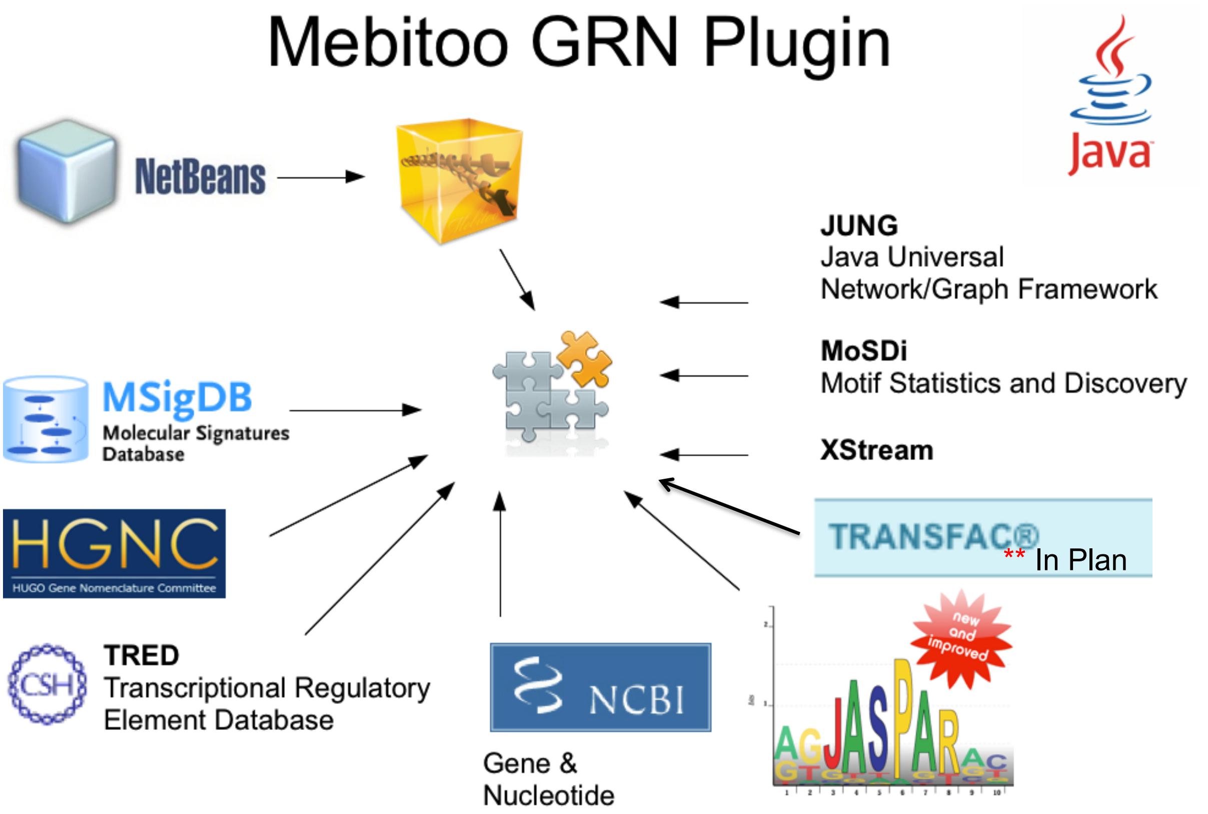
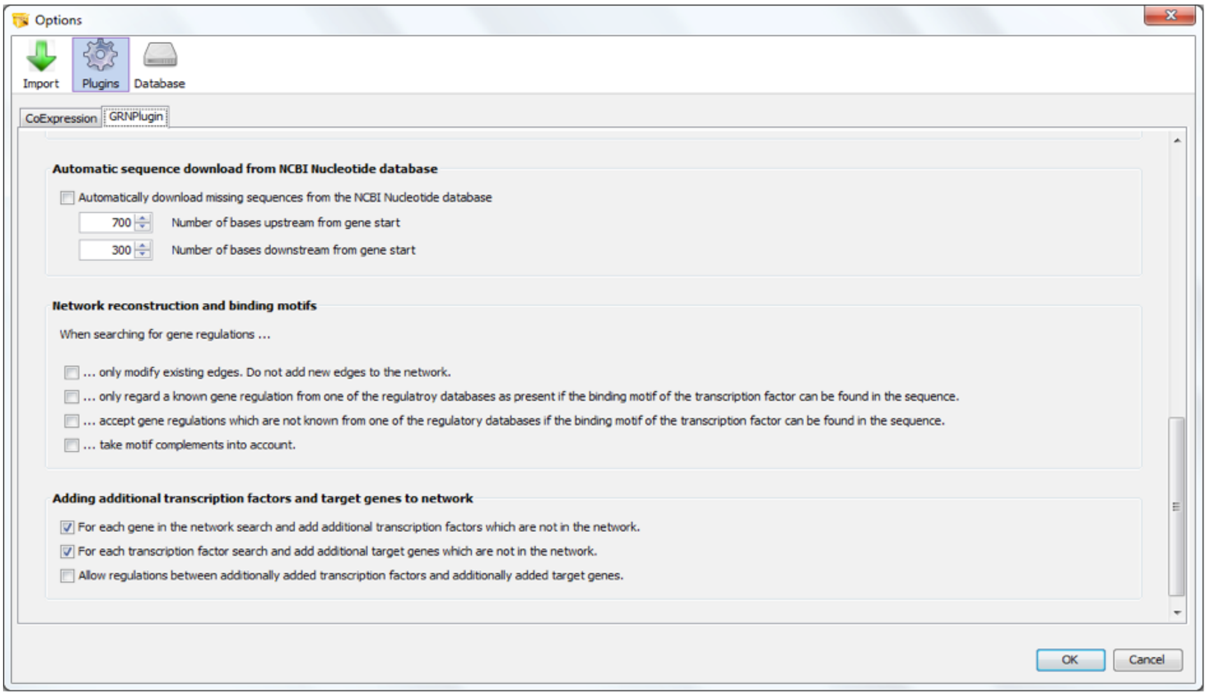


Figure 3-5 The integrated genomic resources and SW components used in the GRN query plugin.
Source: M.Sc. thesis by Johannes Trumm.



Automatic sequence download from NCBI Nucleotide database

□ Automatically download missing sequences from the NCBI Nucleotide database

700 Number of bases upstream from gene start

300 Number of bases downstream from gene start

Network reconstruction and binding motifs

When searching for gene regulations ...

□ ... only modify existing edges. Do not add new edges to the network.

□ ... only regard a known gene regulation from one of the regulatroy databases as present if the binding motif of the transcription factor can be found in the sequence.

□ ... accept gene regulations which are not known from one of the regulatory databases if the binding motif of the transcription factor can be found in the sequence.

□ ... take motif complements into account.

Adding additional transcription factors and target genes to network

☑ For each gene in the network search and add additional transcription factors which are not in the network.

☑ For each transcription factor search and add additional target genes which are not in the network.

□ Allow regulations between additionally added transcription factors and additionally added target genes.

Figure 3-6 User control panel to set the parameters for the GRN query plugin.

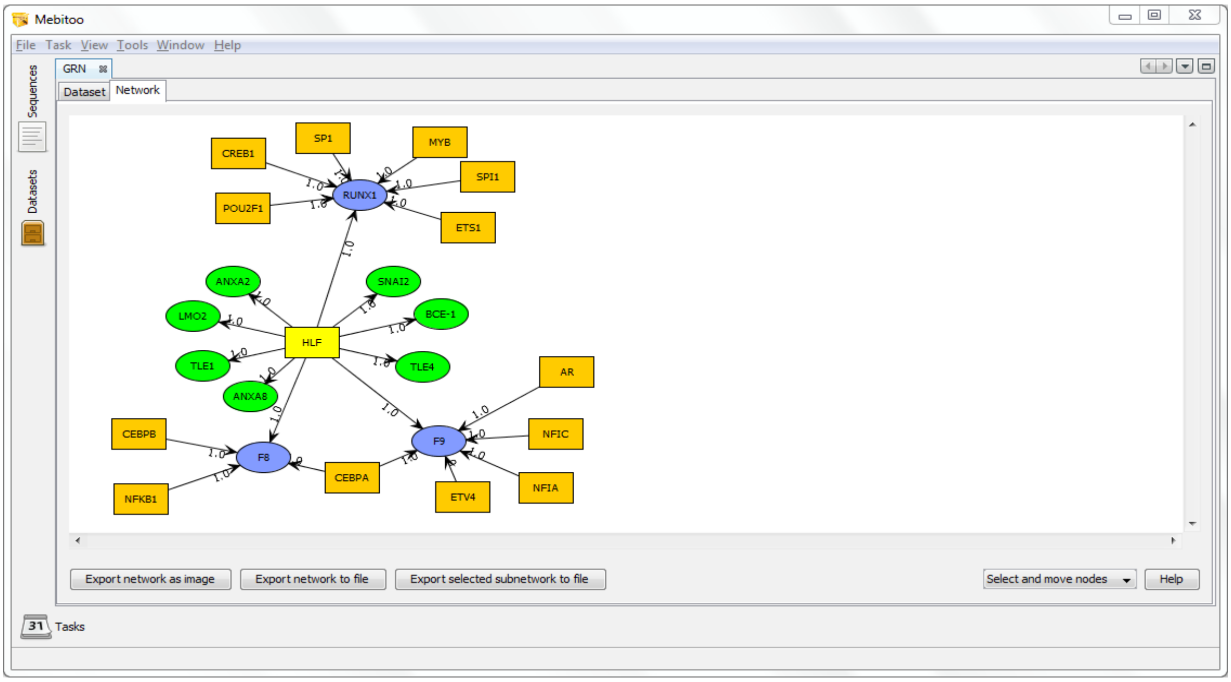


Figure 3-7 Results of the GRN query plugin- Interactive network visualization and export options. Transcription factors involved in the input network are identified and marked in yellow while the remaining genes are colored blue. The tool can expand the input network by adding additional transcription factors (marked in orange) that are annotated as known regulators of the input genes and also by adding additional target genes (marked in green) that are annotated to be regulated by the TFs in the input network.

3.2.3 Plugin 3: learning the network topology using Bayesian network

This plugin is still running as a script and needs to be coupled and integrated to Mebitoo framework. In this step (plugin 3), we constructed a causal probabilistic Bayesian network from the co-expression modules where we used the directed edges obtained from plugin 2 as a start search point to infer directionality between nodes.

As discussed in chapter two, learning of network structures using a Bayesian approach requires a scoring function to assess how well a certain structure fits the input data and a search algorithm to find structures with high scores. Here, we can adopt the greedy algorithm as a search algorithm and the likelihood-equivalence Bayesian Dirichlet (BDe) [87] method as a scoring function for assessing network topology, see chapter 2.

Instead of taking the best network structure that has the best score, we perform the learning approach three times and select only edges that were inferred at least twice in the three runs (edge confidence level ≥ 66.6%).

3.3 Mebitoo: An extensible software framework hosting the pipeline plugins

3.3.1 Description

The Mebitoo framework has mainly been developed by Mr .Christian Spaniol who is another PhD student in the Helms group. During the time line of this PhD thesis, the author was involved in extending some functionalities of Mebitoo and in writing new add-on modules and plugins.

Mebitoo is a software application suite written in Java that is based on the Netbeans Rich-Client platform (RCP) project that can easily be extended with additional functionality deployed as modules. Moreover, the software enables persistent storage with an incorporated database engine, which supports XML files for customized data structures. Since the Mebitoo framework implements a uniform plugin interface, automated data processing can be invoked using a task execution interface in order to queue multiple operations of different modules and process datasets in parallel.

Mebitoo is appropriate for inexperienced users, researchers without programming knowledge as well as scientific programmers, and developers. Aiming at the first group, an easy and friendly GUI is provided that guides the user to sequentially define his tasks (every task represents a one-time running module) and gets the final results in one–click press button. For advanced users with knowledge in Java programming, Mebitoo can be used as a ready hosting workflow automation framework for coupling more new bioinformatics add-on plugins or modules.

3.3.2 Software design consideration

As shown above, the three GRN steps were implemented as independent Java modules/plugins of the Mebitoo framework. Figure 3-8 illustrates the system architecture of the GRN plugins hosted by Mebitoo and the followed design paradigms. The GRN pipeline is designed to support both thin and fat client paradigms. Currently, it works on the fat client paradigm where all business logic and data processing occur in the desktop version installed on the user machine. However, to achieve better performance in data processing, these plugins could be easily switched to the thin client paradigm once the fat server configuration (application server) is available.

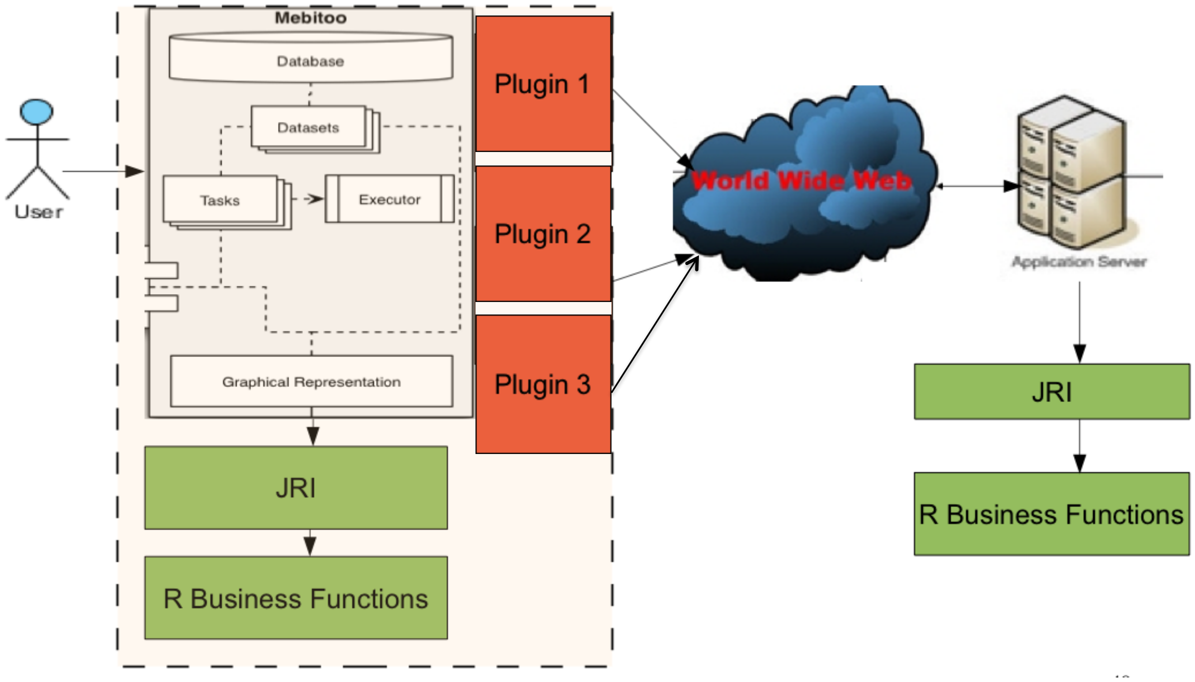


Figure 3-8 SW architecture and design paradigms of the GRN pipeline modules hosted by Mebitoo. Business logic and data processing functions are implemented in R and located on the client version in case of fat client model or located on application server in case of thin client model. The business logic functions are invoked by plugins GUI through the Java–R interfaces JRI.

3.4 Integrative network-based approach

To date, a large number of various methods have been developed to investigate the molecular basis of complex diseases and integrate heterogeneous sources of genomic data. Due to the complexity of the disease pathways and the underlying biology, it is still challenging to integrate and extract meaningful information from large genomic datasets (See literature of chapter 6 for detailed examples). In this regard, we presented a network-based approach [7] utilizing the GRN pipeline explained above to elucidate the regulatory mechanisms of several disease pathways at the molecular transcriptional and post-transcriptional level. Sofar, our approach reports the computational analysis of gene and miRNA transcriptomes, DNA methylome, and somatic mutations to highlight putative disease drivers and novel targets for treatment.

This approach was applied to breast cancer data downloaded from the TCGA portal [66] and was able to reveal strong associations between regulatory elements from different genomic data sources (see chapter 6). The integrated molecular analysis enabled by this approach substantially expands our knowledge base of prospective genomic drivers of genes, miRNAs, and mutations and highlighted candidates for further investigation in the wet lab as novel targets for breast cancer treatment (chapter 6). The provided network-based approach can be applied in a similar fashion to other cancer types, complex diseases, or for studying cellular differentiation processes where such multi-dimensional datasets are available. The integrative network-based approach illustrated in Figure 3-9 currently is able to process four different genomic datasets: gene expression, DNA methylation, miRNA expression, and somatic mutations from normal and diseased cohorts.

3.4.1 Data consistency and preprocessing

For consistency, we considered only samples that were common between all four datasets. For both gene expression and methylation datasets, all probes containing NA values or that were annotated to unknown or multiple genes were removed. Also, probes values were merged by computing the mean of all probes related to single genes within a single sample as previously described in [88].

From the DNA methylation data, we kept only those probes representing CpG sites in the promoter regions of genes. For this, we used the transcription start sites (TSS) for all human genes as annotated in the Eukaryotic Promoter Database EPD [89]. Promoter regions were defined as an interval of ±2kb around the TSS as described in [70]. Then we selected only those CpG sites whose genomic coordinates are contained in that interval.

3.4.2 Differential analysis

The differential expression/methylation analysis was performed using three methods: 1) Significance Analysis of Microarray (SAM) [90], 2) moderated t-test [91], and 3) area under the curve of the receiver operator characteristics (AUC ROC) [91]. Genes that were classified as differentially expressed/methylated genes by at least two of those three methods were included in the list of differentially expressed/methylated genes. The same procedure was applied to determine differentially expressed miRNAs

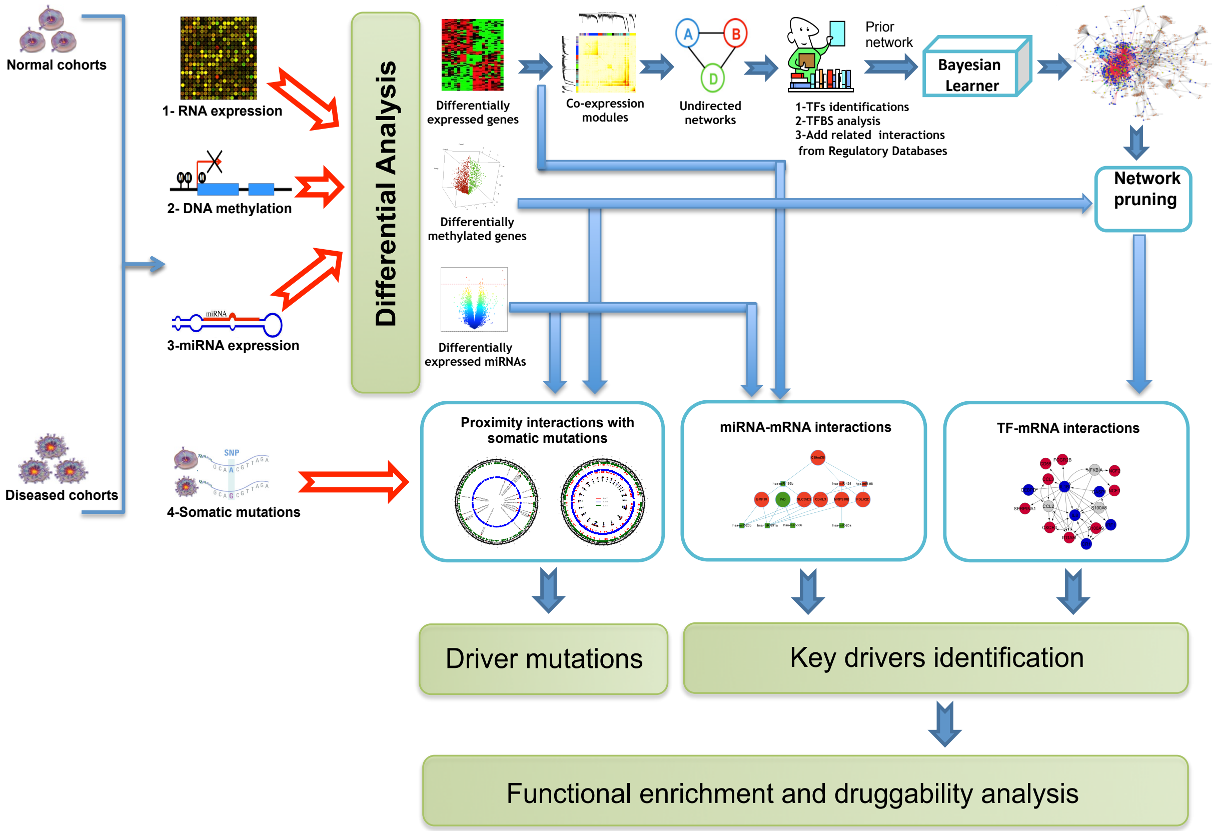


Figure 3-9 the integrative network-based approach.
A schematic diagram describing data processing and integration of different data sources to detect major determinants and key driver molecules.

3.4.3 Network construction using the GRN pipeline

The GRN construction pipeline that consists of plugins 1 to 3 was applied to the differentially expressed genes to obtain the GRN network involving DNA binding proteins (TF) and target genes. Using plugin 1, we constructed from the identified differentially expressed genes the co-expression network based on the pairwise correlation as a distance measure.

The resultant co-expression networks were subjected to plugin 2 as input parameters. Gene interactions suggested from the co-expression networks were connected to regulatory information retrieved from the Transcriptional Regulatory Element Database (TRED) [82], Molecular Signatures Database (MSigDB) [70], and JASPAR database [83]. All genes involved in the co-expression network and listed in at least one of the databases to code for a transcription factor (TF) were marked as TFs. Then, for each TF-gene link in the co-expression network, we searched whether the databases contain a known regulation for this TF-target gene pair. In each of these cases, a directed edge was added between the transcription factor and the target gene. Also, we used the Motif Statistics and Discovery (MoSDi) [84] software to conduct a motif search for all known binding motifs of the TFs represented in the co-expression network against the promoter regions of all genes in the network. If a match was found, a new directed edge from the TF to the gene was added. In the last step (plugin 3), we constructed a causal probabilistic Bayesian network from the co-expression modules where we used the directed edges obtained from plugin 2 as a start search point to infer directionality between nodes, see chapter 2 for more details.

As candidate set of the final directed interactions, we considered directed edges from plugin 2 as well as directed edges confirmed by both plugin 1 and plugin 3. Subsequently, the entire network containing both directed and undirected interactions was exposed to the pruning step explained below. The GRN network was visualized using the igraph [92] package in R as will be illustrated in chapter 6.

3.4.4 Pruning the GRN using methylation and expression profiles

GRN pruning was carried out based on the observation that some genes show increased promoter DNA methylation levels coupled to a remarkable decline of their expression [79]. In such cases, we assumed that the downregulation of gene expression results from the increase in promoter methylation and not due to TFs or miRNAs regulation. Thus, we removed regulatory interactions whose target genes had absolute anti-correlation between their expression and methylation profiles above a selected threshold of 0.65.

3.4.5 Constructing miRNA-mRNA interactions

The integrated association of the differentially expressed miRNAs and the differentially expressed genes (mRNAs) involved three steps. First, for the set of differentially expressed miRNAs, which were either up- or down-regulated between the tumor and normal samples, we used miRTrail [93] via MicroCosm Targets V5 (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/) to extract their target mRNAs (regulated genes) and overlapped them with the identified differentially expressed mRNAs. Second, we used the experimentally validated database TransmiR

[30] to retrieve the regulatory genes (TFs) that potentially regulate the differentially expressed miRNAs. In both steps, the hypergeometric test with a p-value threshold of 0.05 was applied to test the regulation dependencies between the differentially expressed miRNAs and their target genes /their regulatory TFs. Finally, both miRNA→ mRNA (including TF genes) interaction pairs from step one and TF→ miRNA interaction pairs from step two were joined and merged to a final network.

3.4.6 Identifying the genetic key drivers/determinants

Key regulators in the constructed networks were identified by determining the minimal set of nodes that regulate (i.e dominate) the entire network. This problem can be modeled as the following optimization problem:

Let graph *G(V,E)* be a connected graph, *n* = *|V|*, *adj* is the adjacency matrix of *G*, and *adj*(*i*, *i*) = 0, *X* is a binary array of size *n*, such that *X*(*i*) = 1 if node *i* was marked as a key node, and 0 otherwise. Then the objective function is:

$$\min \sum_{i=1}^{n} X(i)$$

$$\text{subject to } \forall i \sum_{i}^{n} adj(i,j).X(j) >= 1.$$

The last constraint guarantees that every node in the network must have at least one key node in its neighborhood. To solve such an optimization problem, we used the linear programming gplk solver [94] via the numerical optimization package OpenOpt [95].

3.4.7 Enrichment and druggability analysis

For gene set enrichment analysis, KEGG pathways and GO functional categories were identified using the DAVID [68] tool. Briefly, we determined which pathways/functional terms were annotated to at least two genes and were statistically overrepresented in the study gene set. Enrichment was evaluated through the hyper-geometric test using a p-value threshold of 0.05 as explained in details in chapter 2. For the enrichment analysis of the miRNAs set, we used the TAM tool [96] which also uses the hyper-geometric test . Druggability analysis of the identified driver genes was performed using the PharmGKB [97], CTD [98], and CancerResource [99] databases.

3.5 SnvDMiR: Associating the genomic proximity of genetic variants with deregulated miRNAs and differentially methylated regions

Although next generation sequencing of diseased traits has unraveled thousands of DNA alterations, the functional relevance of most of these mutations and how they relate to other epigenetic mechanisms are still poorly understood. Alexander Zapp developed in his M.Sc. thesis under the direction of the author of this PhD thesis a small tool SnvDMiR as a freely–available R pipeline that conducts combinatorial proximity analysis between disease–associated SNVs, deregulated miRNAs, and differentially methylated regions (DMRs) to identify genomically adjacent SNV-miRNA pairs as well as SNV-DMR pairs. These variants could be further investigated as putative candidates for driving pathogenic processes in diseases. We demonstrated the usefulness of the SnvDMiR

pipeline by applying it on a published set of breast cancer-related mutations, deregulated miRNAs, and DMRs. Our pipeline characterized potential driver mutations that are predicted to have damaging effects on related protein functions. Availability: http://gepard.bioinformatik.uni-saarland.de/software.
Background

To further our understanding of human oncogenesis, high-throughput sequencing of tumor genomes has uncovered thousands of DNA alterations such as somatic mutations of single nucleotide variants (SNVs) that may be important for tumor initiation or progression [100-106]. Nevertheless, it remains a pressing challenge to determine which mutations are key drivers for tumor pathophysiology and which ones are passengers with no functional effects. To address this need, several approaches have been presented to characterize driver missense mutations [103, 107-109]. Most straightforward is the annotation of non-synonymous mutations in oncogenes or tumor suppressors. In contrast, relatively little attention has been paid to cases where driver mutations could be in close genomic proximity to disease-related genes, miRNAs, or methylated CpG sites.

Chapter 1 explained the importance of DNA methylation and the phenomenon of differential methylation as well as miRNAs and their correlations to genetic mutations. In this regard, the recent availability of disease-related genomic data such as somatic mutations, associated DMRs and miRNAs calls for the development of integrative genomic proximity-based approaches to better understand the functional relevance of most of these mutations and how they relate to epigenetic marks. To this end, we developed SnvDMiR as a freely–available R pipeline that is able to conduct combinatorial proximity analysis between disease–associated SNVs, deregulated miRNAs, and DMRs to identify genomically adjacent SNV-miRNA pairs as well as SNV-DMR pairs. We demonstrated these features on breast cancer-related datasets and the matched SNVs suggested putative driver mutations that could play a critical role in breast cancerogenesis (chapter 6).

3.5.1 Implementation

SnvDMiR is a computational pipeline implemented in R (Figure 3-10). Based on lists of genomic variants, deregulated miRNAs, differentially methylated sites, and user defined parameters (configurations), SnvDMiR investigates whether the significantly deregulated miRNAs and differentially methylated sites are in close genomic vicinity to the provided genomic variants and outputs matching entries in tabular and ideograms plots. The user needs only to run the main script SnvDMiR.R which in turn loads the required libraries/packages, carries out the analysis on the input data, and visualizes the matched entries in genomic ideograms with circular layouts.

For matching miRNAs and somatic variants, the genomic coordinates of the significantly deregulated miRNAs were downloaded from miRBase [110]. Then, SnvDMiR searches for the miRNA sequences in a predefined genomic window (default is 250kb [111]) around each somatic variant. The window size can be set in the configuration file attached with the SnvDMiR script. The matched miRNA-SNV pairs, where the miRNAs occur within the window around the SNV location, are extracted into the som-miRNA-matches.txt file in the output folder.

The second part of the SnvDMiR functionality is to explore whether differentially methylated regions (usually CpG islands) are in the vicinity of somatic mutations. To this end, our tool tests the occurrence of the SNV within a certain genomic distance (default is 3kb) from the genomic coordinates of the differentially methylated sites. The default setting of the predefined distance in the configuration file (3kb) was based on the maximum considered length of typical CpG islands, that is, 500bp [21] ≤ CpG islands ≤ 3kb [112]. Moreover, the user has the option to investigate only the C->A, C->G, and C->T SNVs instead of all mutations via setting the parameter filter_Cytosine in the configuration file. The matched entries are also exported to som-DMR-matches.txt file in the output folder.

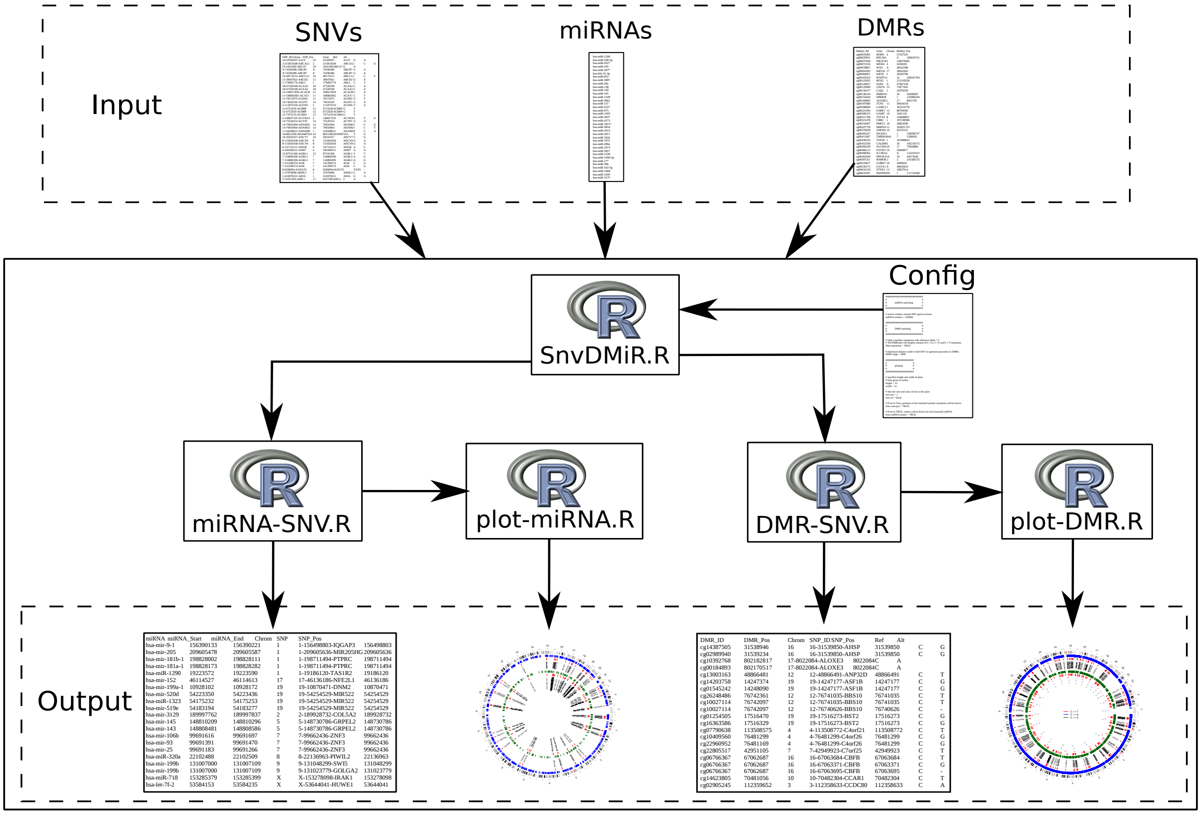


Figure 3-10 The data model for the SnvDmiR proximity pipeline.
The pipeline is used to investigate the vicinity of genetic variants to the deregulated miRNAs and differentially methylated regions. Source: modified from @ Alexander zapp M.Sc. thesis.

Finally, the SnvDmiR utilizes the circlize R package [113] to efficiently plot the related ideogram and flexibly visualize the matched entries in a circular layout as well as the entire input data (all SNVs and either all miRNAs or all DMRs) as genomic background. This helps to better understand the genomic patterns behind the matched entries.

3.6 TFmiR web server

The TFmiR web server was developed in a collaborative fashion of the author of this thesis together with Christian Spaniol and Maryam Nazarieh. The contribution of Maryam was the design and computation of minimum connected dominating sets (see below). The contribution of Christian was the design and implementation of the presentation layer of the web server. The author of this thesis designed and implemented the methodology of setting up integrated regulatory databases, constructing all relative networks, performing statistical analysis as well as the down stream network analysis (as explained below for the breast cancer case study).

We developed TFmiR [114] as a freely available web server for deep and integrative analysis of combinatorial regulatory interactions between TFs/genes and miRNAs that are involved in disease pathogenesis. Since the biological function of molecular components are highly connected with the underlying cellular malfunctions and disease pathways, TFmiR helps to elucidate their cellular mechanisms on the molecular level from a network perspective. The provided topological and functional analyses promote TFmiR as a reliable systems biology tool for researchers across the life science communities. TFmiR web server is accessible through the following URL: http://service.bioinformatik.uni-saarland.de/tfmir.

3.6.1 Background

TFs and miRNAs frequently form Feed Forward Loops (FFLs) and other network motifs to regulate gene transcription in a collaborative manner [115-118]. Therefore, utilizing the combinatorial regulatory information on TFs and miRNAs and their target genes could shed light on key driver genes and miRNAs in human diseases and, in turn, suggests novel therapeutic strategies in disease treatment [7, 115].

Several databases have been developed in order to facilitate the research on transcriptional and posttranscriptional interaction types between TFs/genes and miRNAs. For instance, TransFac [71], OregAnno [119], and MsigDB [70] maintain interactions of TFs regulating genes (TF→gene). TransmiR [30] provides information on which TFs regulate miRNAs (TF→miRNA). mirTarBase[72], TarBase [73] and miRecords [74] collect target genes of miRNAs (miRNA→genes) in different organisms. Although still little is known about miRNA-mediated miRNA regulations, recent studies have reported plausible evidences that miRNAs may regulate the expression of other miRNAs as well as their target genes [120-124]. Thus, miRNA→miRNA interactions were computationally predicted and maintained in the PmmR database [75].

Despite the general availability of such databases, generalized repositories integrating different kinds of molecular interactions and intensively analyzing their contributions to diseases are still missing. To this end, we developed TFmiR, a web server that allows for integrative and comprehensive analysis of interactions between a set of deregulated TFs/genes and a set of deregulated miRNAs within the relevant pathways of a certain disease. It unravels the disease-specific co-regulatory network between TFs and miRNAs and performs over representation analysis (ORA) for the involved TFs/genes and miRNAs. Our web server also detects feed forward loops (FFLs) consisting of miRNAs, TFs, and co-targeted genes (TF-miRNA co-regulatory motifs) and assesses the functional homogeneity between the co-regulated targets in terms of their statistical significance.

Furthermore, TFmiR utilizes 7 different methods for identifying key network players that could possibly drive oncogenic processes of diseases and thus act as potential drug targets. Especially when combined with experimental validation, these putative key players as well as the novel TF-miRNA co-regulatory motifs could promote novel insights to develop new therapeutic approaches for human diseases. Overall, TFmiR presents a comprehensive analysis suite for studying the architecture and feature of the TF-miRNA co-regulatory network.

3.6.2 Description

TFmiR is a freely available web server that integrates genome-wide transcriptional and post-transcriptional regulatory interactions to elucidate human diseases. Based on a selected disease and user-supplied lists of deregulated genes/TFs and miRNAs, TFmiR investigates four different types of interactions, TF→gene, TF→miRNA, miRNA→miRNA, miRNA→gene. It also unravels the interplay circuitry between miRNAs, TFs and target genes within the pathogenicity of the specified disease in a systems biology approach. For each interaction type, TFmiR utilizes information provided by well -established and finely-curated regulatory databases of both predicted and experimentally validated interactions (Figure 3-11) whereby all repeated interactions were removed. For TF→miRNA interactions, we also integrated manually curated regulatory relationships from large numbers (~5000) of published papers (PMID: 20584335) [125]. From the predicted miRNA→miRNA interactions in the PmmR database [75], we considered only the best hits having score < 0.2 which is computed as the normalized path length between the two involved miRNAs. The incorporated predicted miRNA→gene interactions were retrieved from starBase [126] by selecting only those predictions confirmed by three out of five prediction algorithms (targetScan [127], picTar [128],RNA22 [129], PITA [130], and miRanda [131]). Table 3-1 lists the included databases and the number of regulations available for each interaction type. In total, TFmiR integrates information on almost 10.000 genes, 1856 miRNAs, ~ 3000 diseases including subtypes, and more than 111.000 interactions.

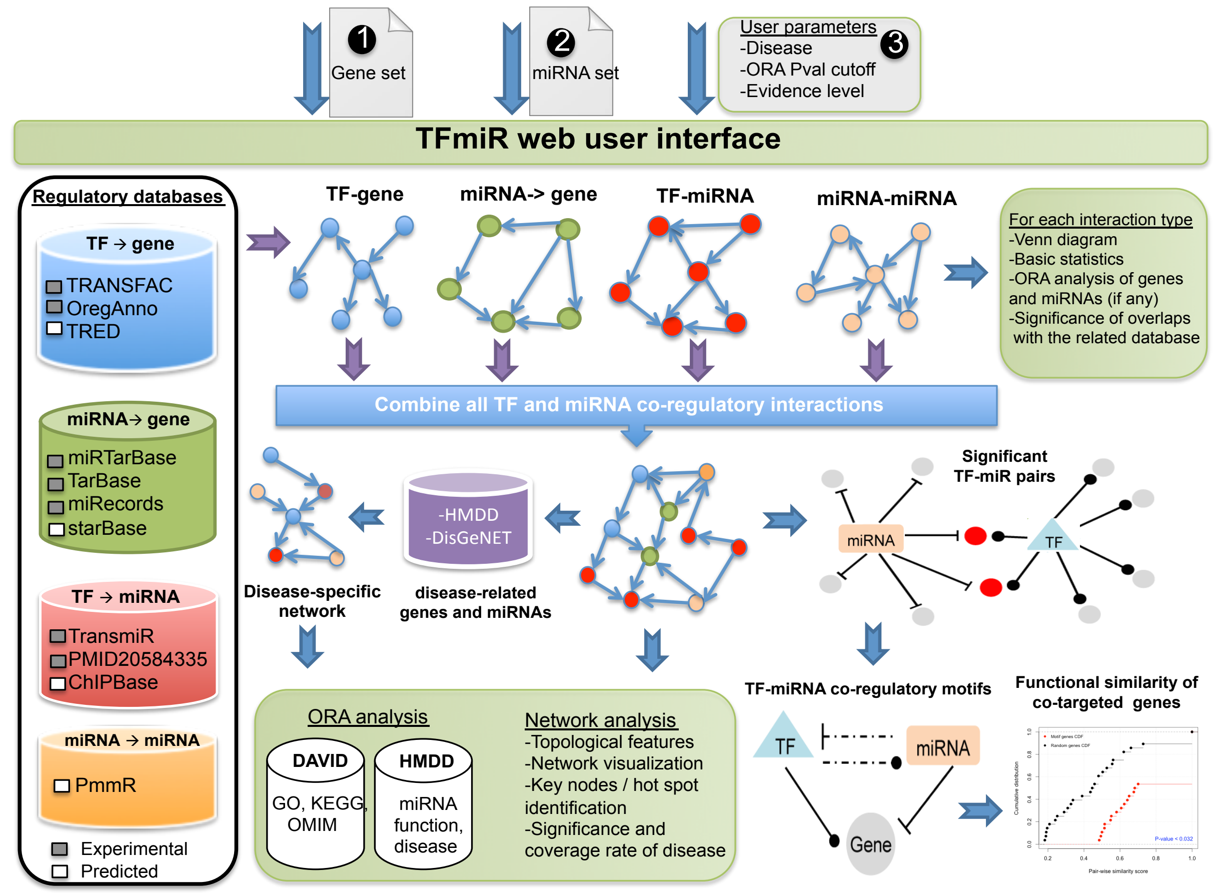


Figure 3-11 A system level overview of the TFmiR architecture.
This schematic diagram describes the incorporated databases, data flows and output downstream analysis.

3.6.3 TFmiR user input scenarios

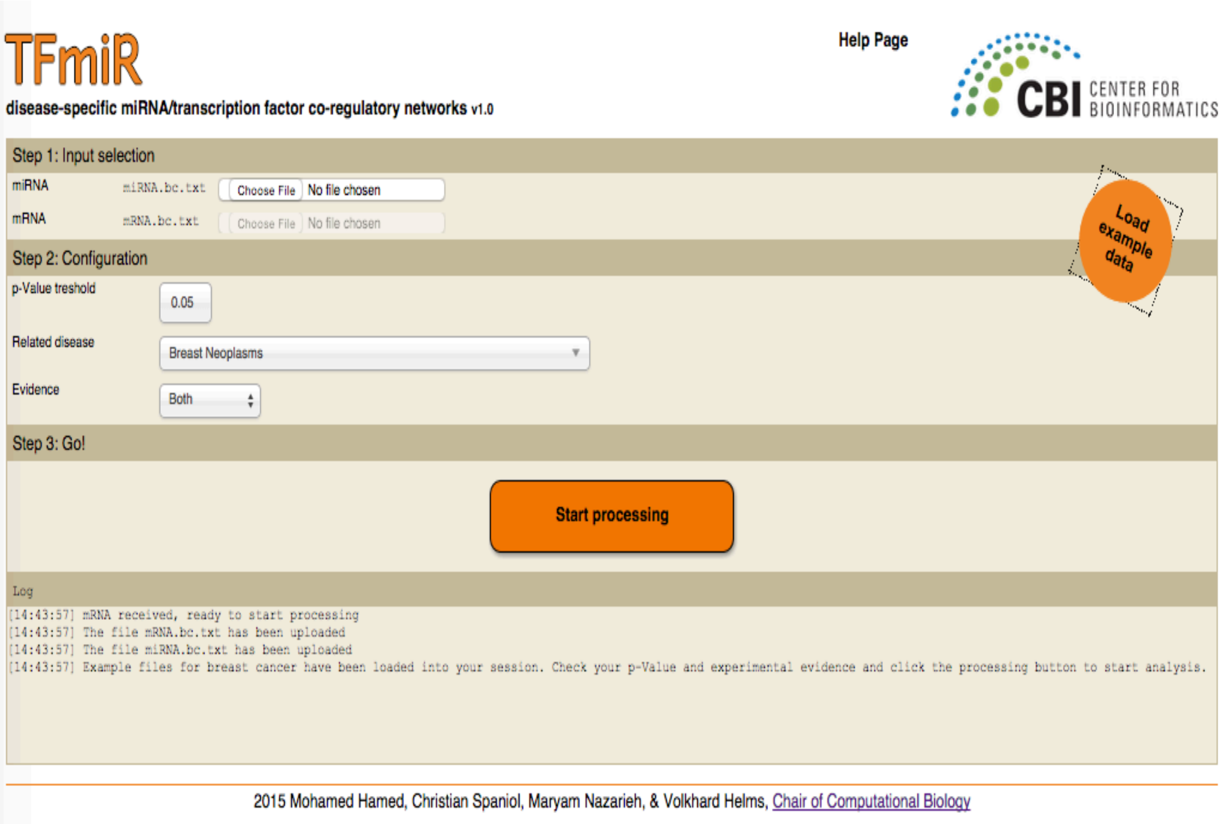
TFmiR can be called through two scenarios. If a user submits two RNA sets (a set of deregulated mRNAs/genes and a set of deregulated miRNAs), the TFmiR web server will return regulatory interactions based on the provided deregulated genes and deregulated miRNAs. In the second scenario, a user submits only a set of deregulated genes. In such a case, TFmiR identifies the set of miRNAs whose target genes as well as regulator TFs are significantly enriched within the input deregulated genes using the hypergeometric distribution function followed by the Benjamini-Hochberg (BH) adjustment with a cutoff value of 0.001. Sample input files of the deregulated genes and miRNAs are provided in the TFmiR web page. The user can optionally set the *p-value* cutoff (default is 0.05) required later for over representation analysis (ORA) on the resulting network nodes (genes / miRNAs), see chapter 2. Finally, the user can control the evidence level (experimentally validated, predicted, or both) for the constructed regulatory interaction that will be subjected later to further network analysis. See Figure 3-12.

Table 3-1 The integrated databases and interaction types in TFmiR.
(P) means predicted interactions and (E) means experimentally validated interactions. All databases were downloaded before August 2014.

Interaction	Databases (P/E) *	Genes	miRNAs	Edges	Version /frozen date
TF→gene	TRANSFAC (E) [71]	1279	--	2943	V11.4
	OregAnno (E)[119]	1132	--	1083	Nov 2010
	TRED (P) [82]	3038	--	6462	2007
TF→miRNA	TransmiR (E) [30]	158	175	567	V1.2, Jan 2013
	PMID20584335 (E) [125]	58	56	102	Apr 2009
	ChipBase (P) [132]	119	1380	33087	V1.1, Nov 2012
miRNA → gene	miRTarBase (E)[72]	2244	551	5640	V4.5, Nov 2013
	TarBase (E) [73]	422	79	492	V7.0
	miRecords (E)[74]	543	157	780	Mar 2009
	starBase (P)[126]	5720	249	56051	V2.0, Sept 2013
miRNA→miRNA	PmmR (P) [75]	--	312	3846	Mar 2011

3.6.4 Functionality of TFmiR

The TFmiR web server pools all the four interactions types based on the significant TF(gene)-miRNA pairs from the input deregulated genes and miRNAs and accordingly generates an entire combinatorial regulatory network, see Figure 3-13. If a disease was selected, TFmiR integrates the human miRNA disease database (HMDD) [133] as well as DisGeNET (a database for gene-disease association) [134] as reliable sources for disease-associated miRNAs and genes, respectively. Interactions whose target nodes or regulator nodes are known to be associated with the disease are composing the putative disease-specific network. As the next step, TFmiR offers a downstream analysis on three different levels: (1) the regulatory subnetwork of each of the four interaction types, (2) the combined network of all interaction types, and (3) the disease-specific network (if disease was selected). For each interaction type subnetwork representing a set of regulator → target links, we display the total number of targets and regulators in the corresponding interaction databases, a Venn diagram depicting the overlap between the input deregulated targets (miRNAs/genes) and the targets of the input deregulated regulators (genes/miRNAs) available from the database. The significance of overlap is computed using the hypergeometric distribution test. To avoid the effect of false-positives in the regulator → target databases and to account for a different number of targets for the input deregulated regulators, a randomization test is conducted (n=1000). Furthermore, the TFmiR web server carries out statistical over representation analysis (ORA) for both gene and miRNA sets comprising the interaction subnetwork.


TFmiR
disease-specific miRNA/transcription factor co-regulatory networks v1.0
Help Page
CBI CENTER FOR BIOINFORMATICS
Choose File
Load example data
Start processing

Step 1: Input selection

miRNA miRNA.bc.txt No file chosen

mRNA mRNA.bc.txt No file chosen

Step 2: Configuration

p-Value treshold 0.05

Related disease Breast Neoplasms

Evidence Both

Step 3: Go!

Log

```
[14:43:57] mRNA received, ready to start processing
[14:43:57] The file mRNA.bc.txt has been uploaded
[14:43:57] The file miRNA.bc.txt has been uploaded
[14:43:57] Example files for breast cancer have been loaded into your session. Check your p-Value and experimental evidence and click the processing button to start analysis.
```

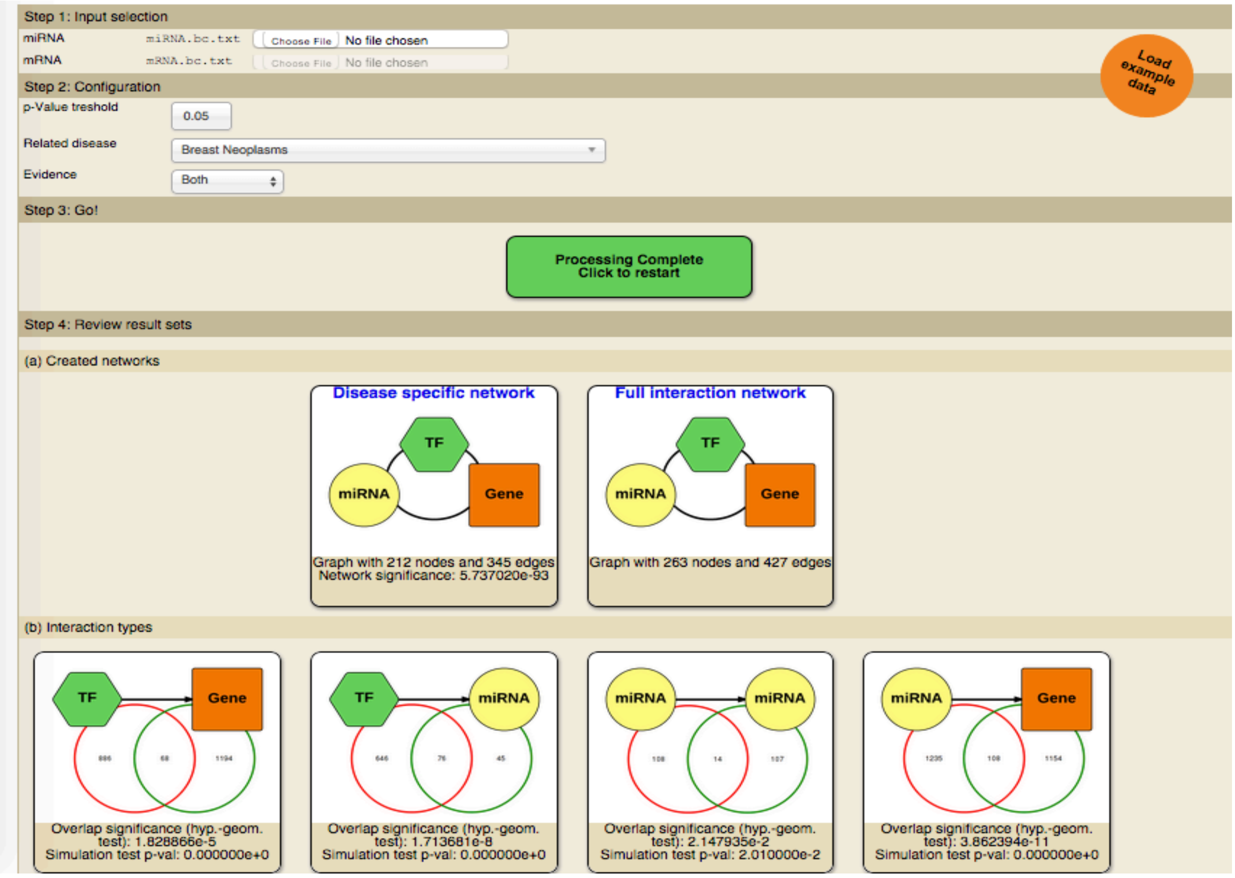
2015 Mohamed Hamed, Christian Spaniol, Maryam Nazarieh, & Volkhard Helms, Chair of Computational Biology

Figure 3-12 TFmiR homepage showing user input parameters.

For gene set analysis, TFmiR employs DAVID [135] to check for enrichment of GO terms (BP subcategory), KEGG pathways, and OMIM diseases as well as a clustering of genes based on their functional similarities. For miRNA set analysis, we used the miRNA-functional association data and miRNA-disease association data from HMDD to statistically relate the functional and disease terms to the miRNA set. For levels 2 and 3, the TFmiR web server calculates for each network the basic topological features, relevance to the disease-associated genes/miRNAs by testing the overlap significance with the network nodes, degree distribution plot, ORA analyses for both gene and miRNA nodes, network key players (hot spots), and detects 3-node motifs. To measure the strength of correlation between the potential disease-specific network, the input disease, and the input deregulated genes and miRNAs, we compute a coverage ratio (*CR*) between the nodes of the disease-specific network and the nodes of the entire combined network.

$$C_R = \frac{N_d}{N_t}$$

Here N_d represents the number of disease-specific network nodes, and N_t represents the total number of nodes in the entire network. We also calculate the *CR* ratio between the edges of the two networks. Along with the aforementioned analysis, all resulting networks are visualized using the interactive Cytoscape-web viewer [136].



Step 1: Input selection

miRNA miRNA.bc.txt Choose File No file chosen

mRNA mRNA.bc.txt Choose File No file chosen

Step 2: Configuration

p-Value treshold 0.05

Related disease Breast Neoplasms

Evidence Both

Step 3: Go!

Processing Complete
Click to restart

Step 4: Review result sets

(a) Created networks

Disease specific network

Graph with 212 nodes and 345 edges
Network significance: 5.737020e-93

Full interaction network

Graph with 263 nodes and 427 edges

(b) Interaction types

Overlap significance (hyp.-geom. test): 1.828866e-5
Simulation test p-val: 0.000000e+0

Overlap significance (hyp.-geom. test): 1.713681e-8
Simulation test p-val: 0.000000e+0

Overlap significance (hyp.-geom. test): 2.147935e-2
Simulation test p-val: 2.010000e-2

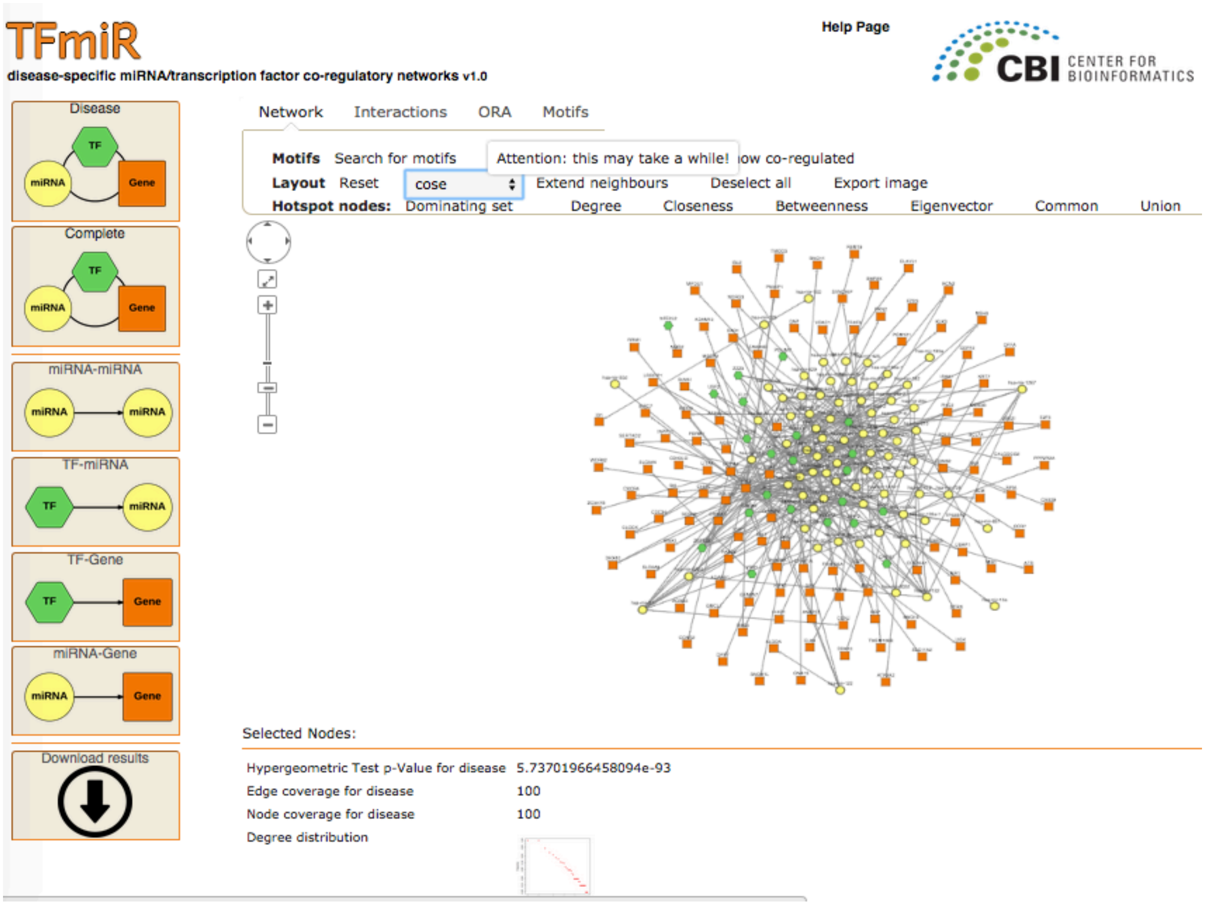
Overlap significance (hyp.-geom. test): 3.862394e-11
Simulation test p-val: 0.000000e+0

Figure 3-13 Reconstructed networks from the input deregulated genes and miRNAs.
Four interaction networks corresponding to the four interactions types as well as the entire interaction network and disease–specific network.

3.6.5 Identification of network key players (hot spots)

To identify crucial network players that could possibly be critical drivers of disease pathogenesis, TFmiR utilizes 7 different methods (Figure 3-14). The first six methods use the well known topological centrality measures: degree centrality, closeness centrality, betweenness centrality, eigenvector centrality as well as the common and union sets of the key nodes identified by these four measures.

We defined the key nodes as the top 10% highest centrality nodes of the TFs, miRNAs, and genes in the disease-specific and whole network. The last method is based on determining the minimal set of nodes that regulate the entire network. We mapped this problem into the Minimum Connected Dominating Set (MCDS) and employed the algorithm presented by Rai et al. 2009 [137] to search for the minimum connected dominating node set. This feature was the contribution of Maryam Nazarieh to the TFmiR publication.



Hypergeometric Test p-Value for disease 5.73701966458094e-93
Edge coverage for disease 100
Node coverage for disease 100

Figure 3-14 Network visualization and key players identification in the TFmiR webserver.

3.6.6 Identification of TF-miRNA co-regulatory motifs

Feed Forward Loops (FFLs) are interconnection patterns that recur in many different parts of a network and form key functional modules [115, 138]. They have been demonstrated as one of the most important motif patterns in transcriptional regulation networks [138] that govern many aspects of normal cell functions and diseases [139-146]. Here, TFmiR identifies 4 types of 3-node motifs (3 FFLs and 1 co-regulation motif) consisting of a TF, a miRNA, and their co-targeted gene that are considered as TF-miRNA co-regulatory motifs (Figure 3-15). (1) The so-called Composite-FFL, which includes TF regulation of both a miRNA and a target gene as well as miRNA suppression of that TF and that target gene. (2) The so-called TF-FFL includes TF regulation of the expression of both a miRNA and a target gene and it also includes miRNA repression of that target gene. (3) The so-called miRNA-FFL includes miRNA repression of both a TF and a target gene, as well as TF regulation of this target gene. (4) The so-called Coregulation-FFL includes only TF regulation of a target gene as well as miRNA repression of that target gene. TFmiR utilizes the following procedure in order to identify the aforementioned motif types.

1-Identifying significant TF-miRNA co-occuring pairs

We identified statistically significant TF and miRNA pairs that cooperatively regulate the same target gene using the hypergeometric distribution and calculated the p-values as given in the following function:

$$P\text{-value} = 1 - \sum_{i=0}^{x} \frac{\binom{k}{i}\binom{M-k}{N-i}}{\binom{M}{N}}$$

where k is the number of target genes regulated by a certain miRNA, N is the number of genes regulated by a certain TF, x is the number of common target genes between this TF and the miRNA, and M is the number of genes in the union of all human genes targeted by human miRNAs combined with all human genes regulated by all human TFs in our databases. Then, a multiple test correction was done by determining the FDR according to the Benjamini and Hochberg (BH) [78] method and only those pairs with a adjusted P-value less than 0.05 were selected as significant TF-miRNA pairs.

2- Construction of candidate TF-miRNA-gene FFLs

All interactions associated with the significant TF-miRNA pairs were represented as connectivity matrix, M, such that $M_{ij} = 1$ if regulator i regulates target j where $i \in$ (TF, miRNA), and $j \in$ (TF, miRNA, gene). Then, we scan all the 3*3 submatrices of M that represent each type of the four considered FFL topologies (Figure 3-15).

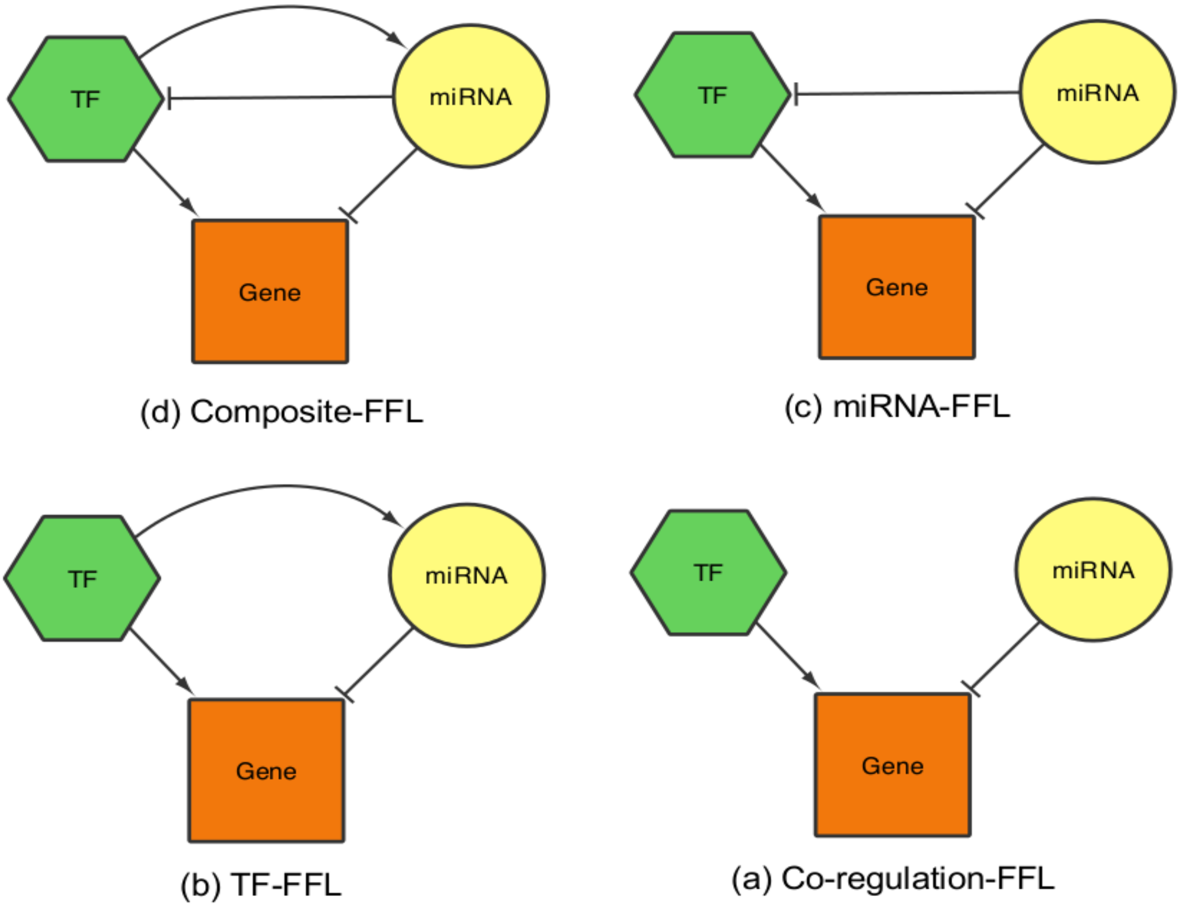


Figure 3-15 Schematic illustration of the four motif types detected in TFmiR.
All motifs contain a TF, a miRNA, and a common target gene.

3-Significance of the FFL motifs

To evaluate the significance of each FFL motif type, we compared the number of times they appear in the real network to the number of times they appear in randomized ensembles preserving the same node degrees. The random networks were constructed 100 times and compared to the real network. A *p-value* is calculated as :

$$P\text{–value} = \frac{\mathrm{N_h}}{\mathrm{N_r}}$$

where N_h is the number of random times that a certain motif type is acquired more than or equal to its number in the real network, and N_r is 100. We also calculate the *Z* score for each motif type to examine by how many standard deviations the observed real motif occurred more often or less often than the mean of the random ones.

$$\mathrm{Zscore} = \frac{\mathrm{N_o} - \mathrm{N_m}}{\sigma}$$

Here N_o is the number of motifs observed in the real network, while N_m, and σ are the mean and standard deviation of the motif occurrence in 100 random networks, respectively.

3.6.7 Functional homogeneity

In order to evaluate the biological evidence of the identified TF-miRNA co-regulatory motifs and better understand their functional roles, TFmiR allows the user to investigate the GO semantic similarity for all pairs of co-targeted genes (genes targeted by the same TF and miRNA pair) or for all pairs of co-regulated genes (all genes regulated by the TF or the miRNAs of that TF-miRNA pair) (Figure 3-16). The GoSemSim R package [147] is used to compute the semantic similarity scores according to the GO annotations. GoSemSim package computes the similarity scores based on the shared GO terms between each pair of genes.

Statistical significance is determined by a permutation test. For this, we randomly select the same number of genes (co-targeted genes or co-regulated genes) from all Entrez genes with Go annotations, and compute their similarity scores. The permutation procedure is repeated 1000 times. Then, we run a Kolmogorov-Smirnov test (KS test) to check whether the functional similarity scores of all gene pairs from the FFL motif are significantly larger than that of randomly selected pairs of genes, see chapter 2 for more details.

3.6.8 Case study

We applied TFmiR to datasets associated with several complex diseases such as cancer, alzheimer and diabetes. In a study on breast cancer (chapter 6), we identified 1262 deregulated genes and 121 deregulated miRNAs using gene and miRNA expression data from the TCGA portal (https://tcga-data.nci.nih.gov/tcga/). These two sets of deregulated genes and miRNAs are in fact the default sample input files now provided with the TFmiR web server. Next, TFmiR was used to reveal the co-regulation network between the deregulated genes/TFs and deregulated miRNAs and to better understand the pathogenic mechanisms associated with breast tumorigenesis. The user input parameters were set as following: *p-value* cut off = 0.05, disease was selected to breast neoplasms, and the evidence level was set to both experimentally validated and predicted interactions.

TFmiR constructed a total of 294 regulatory interactions comprising 172 nodes of deregulated miRNAs and deregulated TFs/genes. The breast cancer-specific network involves 216 interactions and 120 nodes of deregulated miRNAs and genes with node and edge coverage ratios (CR) of 80.6%, and 80.8% respectively. This supports the strong relation between the input deregulated genes and the input deregulated miRNAs in the activity of the oncogenic processes of breast carcinoma. The provided ORA analysis of the disease network nodes reveals their implications in many cancer types as well as cancer-related KEGG pathways. For instance, the network gene nodes are also significantly involved in pancreatic cancer, colorectal cancer, prostate cancer, and the p53 signaling pathway, which is a tumor suppressor gene showing one of the largest frequencies of SNPs among all human genes that have been related to cancer [148]. Moreover, ORA analysis of the network miRNAs shows their involvement in cancerogenesis of multiple organs such as lung neoplasms, ovarian cancer, and adenocarcinoma. Additionally, TFmiR identified 22 key network players (10 genes and 12 miRNAs) based on the union set of four centrality measures described above. These key genes are *E2F6, TP53, SPI1, TGFB1, SMAD4, ESR1, TERT, E2F3, BRCA2, AKT1*, and the key miRNAs are *hsa-mir-148a, hsa-mir-21, hsa-mir-93, hsa-mir-152, hsa-mir-106b, hsa-mir-143, hsa-mir-200c, hsa-mir-27a, hsa-mir-23a, hsa-mir-22, hsa-mir-146a, hsa-mir-335*. Interestingly, some of the identified key genes such as *BRCA2, ESR1, AKT1,* and *TP53* were previously implicated and significantly mutated in breast cancer samples [148]. More importantly, the protein products of the genes *ESR1, TP53, TGFB1, AKT1, and BRCA2* are binding targets for anti-breast cancer drugs [7].

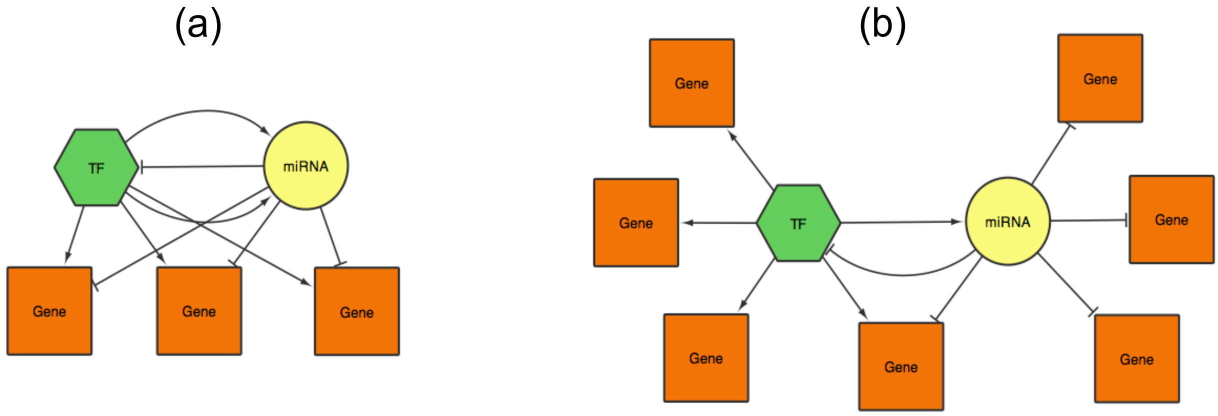


Figure 3-16 Co-targeted and co-regulated genes.
(a) Co-targeted genes defined as genes that are targeted by the same TF and miRNA pair. (b) Co-regulated genes defined as all genes regulated by the TF and the miRNA of this TF-miRNA pair.

It has been demonstrated that the *E2F3* gene plays a critical role in the transcriptional activation of genes that control the rate of proliferation of tumor cells [149-151]. Furthermore, Vimala et al. [152] recently showed that *E2F3* is overexpressed in 11 breast cancer cell lines and *siRNA-E2F3* based gene silencing facilitates the silencing of *E2F3* overexpression and limits the progression of breast tumors. This strongly matches our findings using TFmiR that *E2F3* may be a potential therapeutic target for human breast cancer. The two identified key regulator miRNAs *hsa-mir-143, and hsa-mir-200c* are deregulated tumor suppressor miRNAs in many cancer types [153-155] and are involved in chemotherapy resistance and showed promising insights in the development and delivery of miRNA-based cancer therapeutics [156].

Next, we examined the TF-miRNA co-regulatory motifs that are significantly enriched in the entire interaction network. We identified 53 FFL motifs (3 composite-FFLs, 2 TF-FFLs, 6 miRNA-FFLs, and 42 coreg-FFLs**)**. An interesting motif involves the TF *SPI1,* the miRNA *hsa-mir-155*, and the target gene *FLI1. This* is an example for how FFL motifs hint at better understanding the pathogenicity of breast cancer (Figure 3-17). Recent studies reported that the oncogene SPI1 is involved in tumor progression and metastasis [157-159]. However, the co-regulation of the oncogene *FLI1* [160-162] by both *SPI1* and the oncomiR *hsa-mir-155* was not reported before. However, we show here that the co-regulated target genes of *SPI1* and *hsa-mir-155* have significantly more cellular functions in common than randomly selected genes (Figure 3-18). Hence, this FFL motif provides novel insights on how *SPI1*-and miRNA affect the cellular network in breast cancer and suggests a cooperative functional role between *SPI1* and potential miRNA partners.

In conclusion, unlike the traditional separate analysis of gene expression profiles [163-167] or the aberration of miRNA expression in cancer tissues [168-170], this integrated molecular analysis of deregulated miRNAs and genes using TFmiR was able to uncover important aspects of the TF/gene-miRNA interactomes, their co-regulation mechanisms, and the underlying pathogenesis of human breast cancer

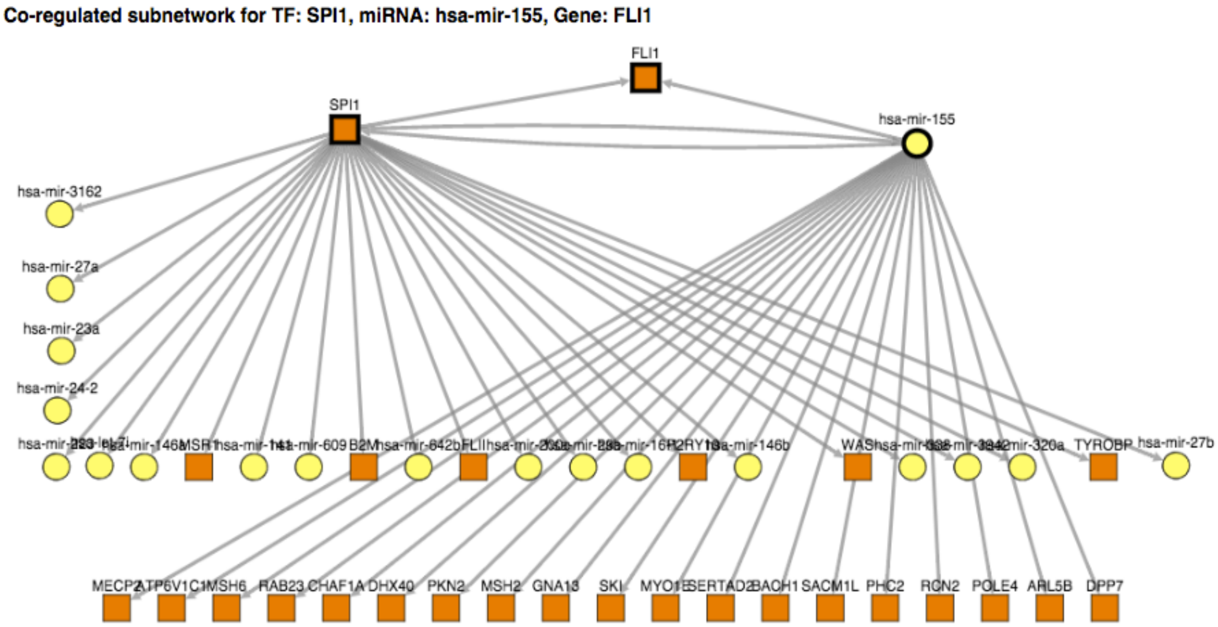


Figure 3-17 A composite FFL motif involves the TF SPI1, the miRNA has-mir-155, and the target gene FLI1. The co-regulated nodes are also visualized and to be further tested for composing a cooperative functional module in breast cancerogenesis.

3.6.9 Comparison with other tools

In comparison with the web interfaces of related databases such as Transmir [30], ChIPBase [132], CircuitsDB [146], starBase [126], and miR2Disease [171], our TFmiR web server has several distinctive features: 1- TFmiR performs integrative analysis of molecular interactions between a set of deregulated genes and a set of deregulated miRNAs within or without the pathogenic pathways of a certain disease. In contrast, the abovementioned web tools can only search the regulatory interactions of a single gene or a single miRNA. 2- TFmiR performs a rich network analysis involving TF-miRNA co-regulatory motif detection, plausible network visualization, statistical significance of the extracted interactions, and ORA analysis for each interaction type, the combined interaction network, and the disease network. Such an integrated analysis is not provided by other web tools. 3- TFmiR allows the user to retrieve either experimentally validated or predicted interactions or both. Such an option is not available using the other tools. In a relatively similar fashion, DisTMGneT [172] was developed for obtaining cancer-specific network based on user-selected sets of deregulated genes and miRNAs. However, it lacks the downstream analysis, the varieties of user input parameters, and it is limited to a predefined set of miRNAs and genes as well as cancer disease. Also miRTrail [93] performs ORA and Gene Set Enrichment (GSEA) analyses of interactions of genes and miRNAs based on expression profiles. However, it explores only miRNA→gene interactions.

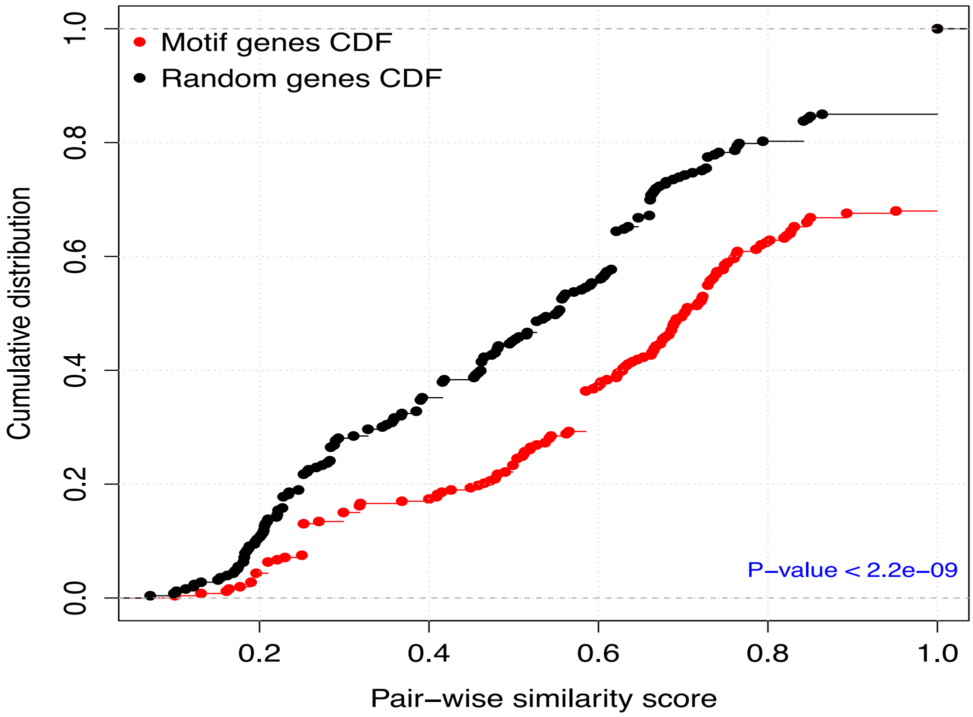


Figure 3-18 Cumulative distributions of GO functional semantic scores of gene pairs of co-regulated genes in the examined motif (red) versus randomly selected genes (black). The p-value was calculated using the Kolmogorov-Smirnov test.

3.6.10 Conclusion

We developed TFmiR as a comprehensive web server for integrative analysis of the molecular interactions between TFs/genes and miRNAs and their interwoven critical roles in the pathology of human diseases. TFmiR shows advances over other related web tools in terms of the extended downstream analysis, the varieties of user parameters, use case scenarios, and in incorporating information from various well-established regulatory databases. TFmiR is based on user-provided sets of deregulated genes and/or miRNAs regardless of the data producing technologies of either microarray experiments, NGS, or PCR. The application of TFmiR on breast cancer–related deregulated genes and mirNAs demonstrated the usefulness of TFmiR in constructing the breast cancer–specific network and identifying literature-confirmed core regulators as well as novel hub nodes of TFs/miRNAs that could be further experimentally investigated as new potential drug targets. TFmiR was also able to characterize important TF–miRNA co-regulatory motifs whose co-regulated genes form cooperative functional modules in breast cancerogenesis.

3.6.11 Outlook

Besides the involved transcriptional and posttranscriptional regulatory interactions, possible extensions are to integrate data for posttranslational events such as protein phosphorylation and localization. Also enriching TFmiR with additional well-established databases and extending the downstream analysis of the interaction networks would be a valuable asset. Furthermore, an extra analysis module of detecting 4-nodes FFL motifs between TFs, miRNAs, and target genes can be coupled into TFmiR. Finally, expanding the TFmiR to elucidate the regulatory mechanisms of cellular processes (ex. stem cell differentiation) in addition to diseases would make TFmiR of great interest to a wide range of researchers in the life science community.

3.7 NGS pipeline

3.7.1 Background

In collaboration with Dr. Ulrich Nübel at the Robert-Koch-Institute institute, we developed a Whole Genome Sequencing (WGS) pipeline to identify core-genome SNPs that can be effectively used to study the phylogenetic arrangements between bacterial isolates as well as an additional module to elucidate phenotypic characteristics such as virulence (Figure 3-19). The pipeline was written in a combination of shell scripting and R language[173].

In a collaboration with Prof. Dr. Lutz von Müller and Prof. Dr. Mathias Herrmann (both medical faculty, Saarland University) and Dr. Patrick Nitsche (HZI Braunschweig), this pipeline was applied to Methicillin-Resistant Staphylococcus Aureus (MRSA) genomes to investigate the phylogenetic positions of the recently emerged t504 clone (Spa-type t504) in the Saarland province of Germany in relation to the currently dominant clone t003 in the surrounding areas of south Germany and Luxemburg. Following this, we analyzed the differentially occurring genetic mutations between nasal and blood stream (invasive) samples of the predominant CC5 with the aim of better understanding the infectivity mechanism of the invasive group. The whole study is introduced in full detail in Chapter 8.

3.7.2 Pipeline description

Whole genome sequencing of MRSA DNA was performed using an Illumina MiSeq sequencer at the HZI in Braunschweig, Germany, producing paired-end reads of 251 basepair lengths with an average coverage of 110-fold.

The first step in the pipeline is quality control of the input sequencing data. For this, the pipeline utilizes the FastQC [174] tool to evaluate the efficacy of the short read data to be involved in the analysis. Secondly reads are mapped against the complete reference genome of interest using the short read alignment version of the Burrows-wheeler Aligner (BWA) algorithm [175]. In our case study presented in chapter 8, we used *S. aureus* CC5 strain NC_017340.1 (http://www.ncbi.nlm.nih.gov/nuccore/NC_017340) as a reference genome. Once the short reads are mapped to the reference genome, we reprocess the mapped reads and investigate the mapping quality distribution (Figure 3-20) such that both duplicate reads and reads with low mapping quality (< 30) are filtered out and the final alignments are sorted via samtools [176].

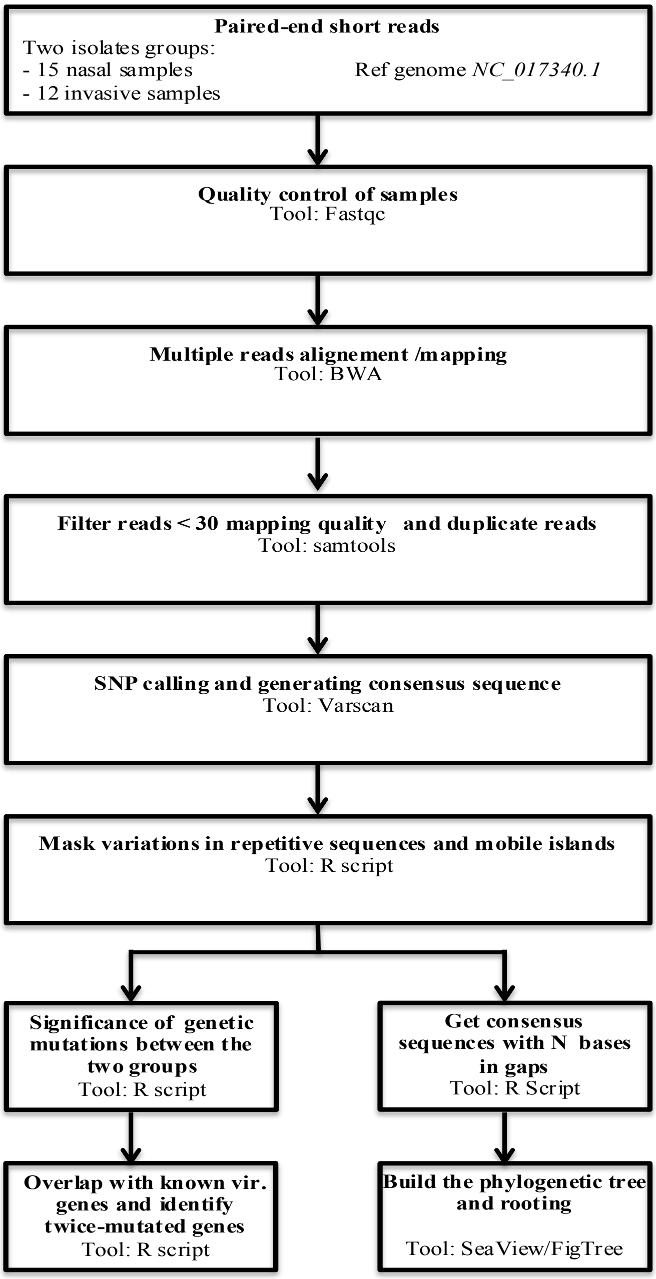

Paired-end short reads
Two isolates groups:
- 15 nasal samples
- 12 invasive samples
Ref genome *NC_017340.1*
Quality control of samples
Tool: Fastqc
Multiple reads alignement /mapping
Tool: BWA
Filter reads < 30 mapping quality and duplicate reads
Tool: samtools
SNP calling and generating consensus sequence
Tool: Varscan
Mask variations in repetitive sequences and mobile islands
Tool: R script
Significance of genetic mutations between the two groups
Tool: R script
Get consensus sequences with N bases in gaps
Tool: R Script
Overlap with known vir. genes and identify twice-mutated genes
Tool: R script
Build the phylogenetic tree and rooting
Tool: SeaView/FigTree

Figure 3-19 NGS pipeline for identifying core-genome SNPs.
And detecting genetic differences between two sets of isolates, such as groups of invasive and nasal MRSA isolates.

As a next step, genetic mutations of the consensus genotypes (SNPs, and INDELS) are called using the VarScan2 [177] tool based on the number of aligned reads supporting each allele.

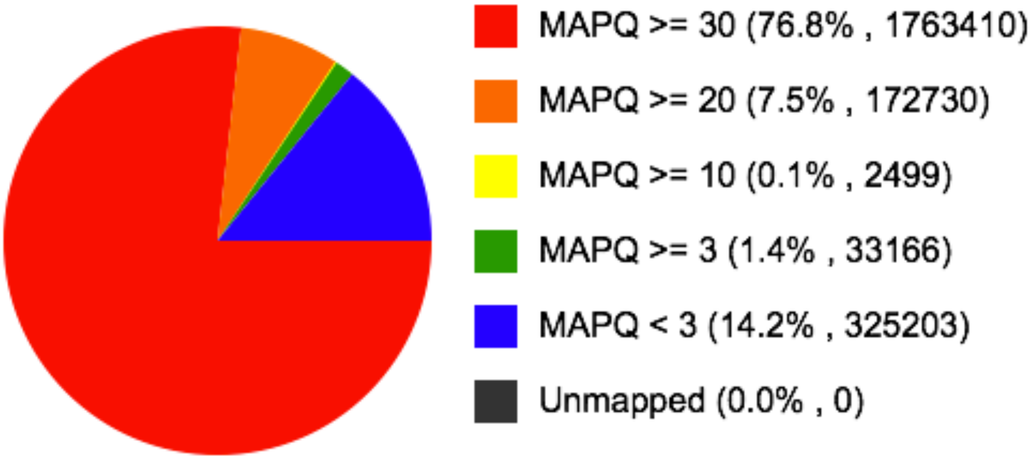


Figure 3-20 An example for mapping quality distribution after the alignment step.

To avoid false positives (phylogenetic mispositioning), the analysis was restricted to the consensus sequence of the highly conserved core-genome. Therefore, variants that occurred in mobile genetic islands as well as repetitive sequence regions were masked by the same reference nucleotide (Figure 3-21). This step was performed using a list of fast evolving regions assembled in the group of Dr. Ulrich Nübel. The reason for this is that these specific genomic regions do evolve randomly in various rates within different strains.

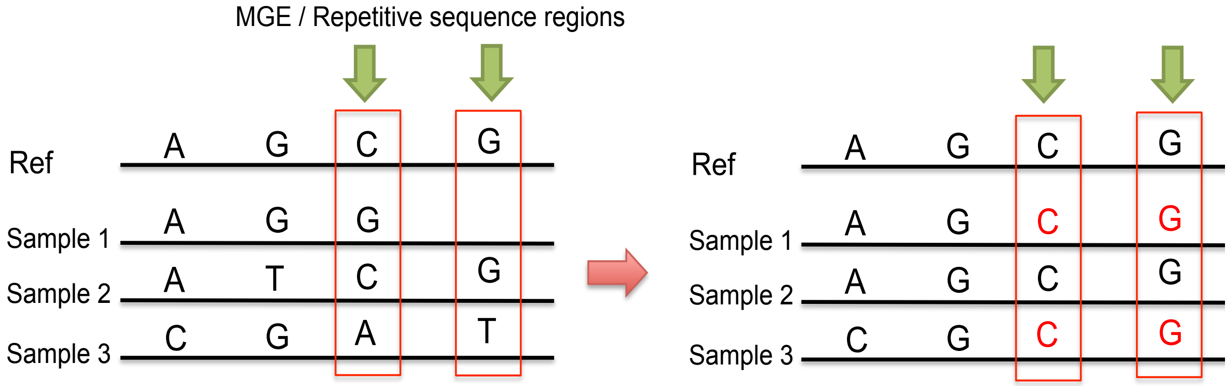


Figure 3-21 Masking genetic variants that occurred in mobile genetic islands or repetitive sequence regions.

Next, a consensus sequence with the same length of the reference genome is constructed for each isolate by padding N nucleotides in the unmapped regions (Figure 3-22).

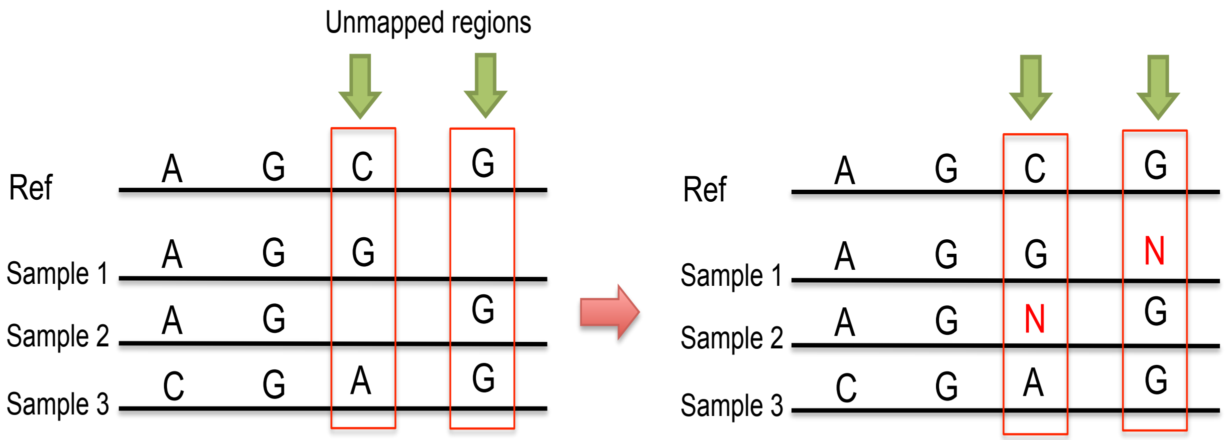
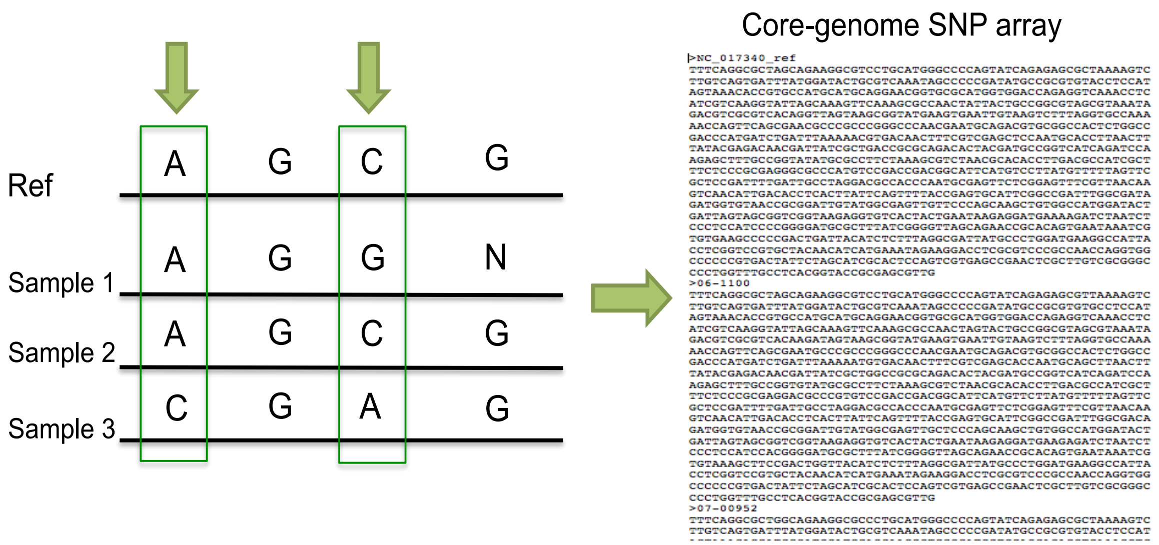


Figure 3-22 Padding the unmapped regions with an N nucleotide to construct the consensus sequence.

In order to obtain a phylogenetic representative core-genome SNP matrix, we considered the genomic positions where at least one variant was found in any of the bacterial isolates (Figure 3-23).



```
>NC_017340_ref
TTTCAGGCGCTAGCAGAAGGCGTCCTGCATGGGCCCCAGTATCAGAGAGCGCTAAAAGTC
TTGTCAGTGATTTATGGATACTGCGTCAAATAGCCCCCGATATGCCGCGTGTACCTCCAT
AGTAAACACCGTGCCATGCATGCAGGAACGGTGCGCATGGTGGACCAGAGGTCAAACCTC
ATCGTCAAGGTATTAGCAAAGTTCAAAGCGCCAACTATTACTGCCGGCGTAGCGTAAATA
GACGTCGCGTCACAGGTTAGTAAGCGGTATGAAGTGAATTGTAAGTCTTTAGGTGCCAAA
AACCAGTTCAGCGAACGCCCGCCCGGGCCCAACGAATGCAGACGTGCGGCCACTCTGGCC
GACCCATGATCTGATTTAAAAACGTGACAACTTTCGTCGAGCTCCAATGCACCTTAACTT
TATACGAGACAACGATTATCGCTGACCGCGCAGACACTACGATGCCGGTCATCAGATCCA
AGAGCTTTGCCGGTATATGCGCCTTCTAAAGCGTCTAACGCACACCTTGACGCCATCGCT
TTCTCCCGCGAGGGCGCCCATGTCCGACCGACGGCATTCATGTCCTTATGTTTTTAGTTC
GCTCCGATTTTGATTGCCTAGGACGCCACCCAATGCGAGTTCTCGGAGTTTCGTTAACAA
GTCAACATTGACACCTCACTTATTCAGTTTTACCGAGTGCATTCGGCCGATTTGGCGATA
GATGGTGTAACCGCGGATTGTATGGCGAGTTGTTCCCAGCAAGCTGTGGCCATGGATACT
GATTAGTAGCGGTCGGTAAGAGGTGTCACTACTGAATAAGAGGATGAAAAGATCTAATCT
CCCTCCATCCCCGGGGATGCGCTTTATCGGGGTTAGCAGAACCGCACAGTGAATAAATCG
TGTGAAGCCCCCGACTGATTACATCTCTTTAGGCGATTATGCCCTGGATGAAGGCCATTA
CCTCGGTCCGTGCTACAACATCATGAAATAGAAGGACCTCGCGTCCCGCCAACCAGGTGG
CCCCCCGTGACTATTCTAGCATCGCACTCCAGTCGTGAGCCGAACTCGCTTGTCGCGGGC
CCCTGGTTTGCCTCACGGTACCGCGAGCGTTG
>06-1100
TTTCAGGCGCTAGCAGAAGGCGTCCTGCATGGGCCCCAGTATCAGAGAGCGTTAAAAGTC
TTGTCAGTGATTTATGGATACTGCGTCAAATAGCCCCCGATATGCCGCGTGTGCCTCCAT
AGTAAACACCGTGCCATGCATGCAGGAACGGTGCGCATGGTGGACCAGAGGTCAAACCTC
ATCGTCAAGGTATTAGCAAAGTTCAAAGCGCCAACTAGTACTGCCGGCGTAGCGTAAATA
GACGTCGCGTCACAAGATAGTAAGCGGTATGAAGTGAATTGTAAGTCTTTAGGTGCCAAA
AACCAGTTCAGCGAATGCCCGCCCGGGCCCAACGAATGCAGACGTGCGGCCACTCTGGCC
GACCCATGATCTGATTTAAAAATGTGACAACTTTCGTCGAGCACCAATGCAGCTTAACTT
TATACGAGACAACGATTATCGCTGGCCGCGCAGACACTACGATGCCGGTCATCAGATCCA
AGAGCTTTGCCGGTGTATGCGCCTTCTAAAGCGTCTAACGCACACCTTGACGCCATCGCT
TTCTCCCGCGAGGACGCCCGTGTCCGACCGACGGCATTCATGTTCTTATGTTTTTAGTTC
GCTCCGATTTTGATTGCCTAGGACGCCACCCAATGCGAGTTCTCGGAGTTTCGTTAACAA
GTCAACATTGACACCTCACTTATTCAGTTTTACCGAGTGCATTCGGCCGATTTGGCGACA
GATGGTGTAACCGCGGATTGTATGGCGAGTTGCTCCCAGCAAGCTGTGGCCATGGATACT
GATTAGTAGCGGTCGGTAAGAGGTGTCACTACTGAATAAGAGGATGAAGAGATCTAATCT
CCCTCCATCCACGGGGATGCGCTTTATCGGGGTTAGCAGAACCGCACAGTGAATAAATCG
TGTAAAGCTTCCGACTGGTTACATCTCTTTAGGCGATTATGCCCTGGATGAAGGCCATTA
CCTCGGTCCGTGCTACAACATCATGAAATAGAAGGACCTCGCGTCCCGCCAACCAGGTGG
CCCCCCGTGACTATTCTAGCATCGCACTCCAGTCGTGAGCCGAACTCGCTTGTCGCGGGC
CCCTGGTTTGCCTCACGGTACCGCGAGCGTTG
>07-00952
TTTCAGGCGCTGGCAGAAGGCGCCCTGCATGGGCCCCAGTATCAGAGAGCGCTAAAAGTC
TTGTCAGTGATTTATGGATACTGCGTCAAATAGCCCCCGATATGCCGCGTGTACCTCCAT
```

Figure 3-23 Constructing the core-genome SNP matrix from the consensus sequences.

Core-genome SNPs from coding and non-coding genomic regions were used to generate a phylogenetic tree using the Maximum Likelihood method as implemented in the SeaView tool [178]. This tree then can be displayed and annotated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

The pipeline identifies the genetic variations between each isolate pair (invasive and nasal) in a similar way as the somatic mutations found between the healthy and disease cohorts. The significance of the acquired genetic variations was evaluated by VarScan on the basis of the sequence reads through Fisher's exact test using a significance level or p-value threshold of 0.05. Successfully passed variants were collected and annotated to the corresponding genes in the reference genome. Subsequently, the variants were grouped by position, and the occurrence of each variant was noted.

4. Imprinted genes and cell differentiation

This chapter is a shortened version of the following publication:

- Mohamed Hamed, Siba Ismael, Martina Paulsen, and Volkhard Helms, Cellular functions of genetically imprinted genes in human and mouse as annotated in the Gene Ontology. PLoS One, 2012. 7(11): p. e50285.

Synopsis

Genomic imprinting is an epigenetic phenomenon that is closely associated with cell development and cellular differentiation. In order to characterize the role of imprinted genes during differentiation processes, the study presented in this chapter was set out to comprehensively investigate the cellular functions of the whole set of imprinted genes, paternally expressed genes, and maternally expressed genes in both human and mouse. Additionally, we examined the transcription factors that are predicted to regulate the imprinted genes and their relatedness to cell differentiation. The findings of this chapter raised intriguing questions regarding the nature and extent of the role of imprinted genes in hematopoietic stem cell differentiation, which will be covered in chapter 5.

Abstract

By analyzing the cellular functions of genetically imprinted genes as annotated in the Gene Ontology for human and mouse, we found that imprinted genes are often involved in developmental, transport and regulatory processes. In human, paternally expressed genes are enriched in GO terms related to the development of organs and of anatomical structures. In mouse, maternally expressed genes regulate cation transport as well as G-protein signaling processes. We noticed that the Gene Ontology currently only provides a partial compilation of which genes are known to be genetically imprinted and what their functions are. Furthermore, we investigated if imprinted genes are regulated by common transcription factors. We identified 25 TF families that showed an enrichment of binding sites in the set of imprinted genes in human and 40 TF families in mouse. In general, maternally and paternally expressed genes are not regulated by different transcription factors. The genes *Nnat, Klf14, Blcap, Gnas* and *Ube3a* contribute most to the enrichment of TF families. In mouse, genes that are maternally expressed in placenta are enriched for AP1 binding sites. In human, we found that these genes possessed binding sites for both, AP1 and SP1.

4.1 Background

Genomic imprinting is an epigenetic phenomenon observed in eutherian mammals. For the large majority of autosomal genes, the two parental copies are both either transcribed or silent. However, in a small group of genes one copy is turned off in a parent-of-origin specific manner thereby resulting in monoallelic expression. These genes are called 'imprinted' because the silenced copy of the gene is epigenetically marked or imprinted in either the egg or the sperm [179].

Imprinted genes play important roles in development and growth both pre- and postnatally by acting in fetal and placental tissues [180]. Interestingly, there appears to exist a general pattern whereby maternally expressed genes tend to limit embryonic growth and paternally expressed genes tend to promote growth. A model case for this striking scenario is the antagonistic action of *Igf2* and *Igf2r* in mouse. Deletion of the paternally expressed *Igf2* gene results in intrauterine growth restriction. On the other hand, deletion of the maternally expressed gene *Igf2r*, results in overgrowth [181].

The observation that maternally and paternally expressed genes apparently act as antagonists has inspired several evolutionary theories that aim to explain the origin of genetic imprinting under the process of ‘natural selection’ [180]. The most scientifically accepted theory is currently the kinship theory [182] and [183]. Briefly, this theory suggests that in polygamous mammalian species, silencing of maternally derived growth inhibiting genes results in increased growth of the embryo. This is associated with an increased nutritional demand and thereby with an exploitation of maternal resources at the cost of future off-spring that might be fathered by a different male.

The evolution of a gene regulatory mechanism that silences preferentially one parental allele of a gene implies that paternally and maternally expressed genes experience different selective pressures during evolution. This assumption is supported by the finding that the two groups reveal different patterns of sequence conservation. Whereas the protein-encoding DNA sequences of paternally expressed genes are well conserved among different mammalian species, maternally expressed genes are much more divergent [184]. Whether paternally and maternally expressed genes differ also in molecular functions and gene regulation is a question that has not yet been investigated in detail.

As the phenomenon of genomic imprinting is an important evolutionary facet of mammals with placentas, it is of great interest to identify which sorts of cellular and developmental processes of developing and/or mature organisms are subject to control by imprinted genes. We aimed in this study at characterizing the cellular roles of imprinted genes in an unbiased, data-driven approach. For this, we used the gene annotations of the Gene Ontology (GO) that consists of three structured and controlled vocabularies for the biological processes, cellular components, and molecular functions associated with particular genes. As it is of particular interest to analyze which of these functions are controlled by the sets of maternally and paternally expressed genes, we have also separately analyzed the enrichment of GO terms in these two groups.

4.2 Methods

4.2.1 Gene Selection

Imprinted genes of human and mouse were downloaded from the Catalogue of Imprinted Genes and Parent-of-origin Effects in Humans and Animals (IGC) [180, 185]. The catalogue encompasses genes that were described as being imprinted in literature. As the related experiments were done in many different labs, the experimental procedures differed considerably. After reading the original publications, we manually selected 64 imprinted genes that are imprinted without doubt in at least one of the two species, see table A-1. This list was provided to us by our collaborator Dr.Martina Paulsen. For the gene *C15orf2*, the expressed allele is unknown since there is no information on the parental origin of the alleles. *Copg2,* and *Zim2* are paternally expressed in the human, but maternally expressed in the mouse. *Grb10* exhibits isoform-specific imprinting effects, i.e. there are paternally expressed and maternally expressed isoforms. The other 60 genes have been experimentally classified into paternally and maternally expressed alleles in two equal halves. 25 genes are imprinted in both species, for the remaining imprinted expression was proven only for one of the two species. As control group for the human (mouse) imprinted genes we used all human (mouse) genes that are annotated in the Gene Ontology.

4.2.2 Functional Enrichment Analysis

For analyzing significantly enriched functional categories, we used the functional annotation tool available in the Database for Annotation, Visualization and Integrated Discovery (DAVID) [135]. We determined which GO categories are statistically overrepresented in different sets of genes. Enrichment was evaluated through the Fisher Exact test using a significance level or p-value threshold of 0.05. We suspected that some functional categories with a high statistical significance may show over-representation even when annotated only to a single gene. In that case, it would not be clear if this function is related to monoallelic expression of the gene in certain tissues, or when it is biallelically expressed in other tissues. Therefore we required that each GO term considered here is annotated to at least two human (mouse) genes.

For the most specific GO terms, we ran the same enrichment analysis procedures by using the biological process GO_FAT database instead of using the general GO knowledgebase. The map enrichment plugin in Cytoscape [85] was used to visualize the overrepresented functional terms and display the overlapping functional sets.

4.2.3 Gene Functional clustering

Clustering and grouping of the imprinted genes were performed using the DAVID gene functional classification tool. This tool employs a set of fuzzy clustering techniques to classify input genes into functionally related gene groups (or classes). This is done on the basis of the co-occurrence of annotation terms by generating a gene-to-gene similarity matrix based on shared functional annotation. This switches the functional annotation analysis from a gene-centric analysis to a biological module-centric analysis [135]. The similarity threshold was set to the minimum similarity threshold of 0.3 suggested by the DAVID consortium. This is then the minimum value to be considered by the similarity-matching algorithm as biologically significant. Also, we set the minimum gene number in a seeding group to 2. This would be the minimum size of each cluster in the final results. All remaining parameters were kept to their default values. The results of the functional classification tool are visualized as heat maps to show the corresponding gene-annotation association across the clustered genes.

4.2.4 Transcription Factor Target Enrichment

The web-based gene set analysis toolkit WebGestalt [77] was used to analyze the targets of transcription factors (TFs). This tool incorporates information from different public resources such as NCBI Gene, GO, KEGG and MsigDB (http://bioinfo.vanderbilt.edu/webgestalt/). Using the TF target analysis tool implemented in WebGestalt, we analyzed whether a set of genes is significantly enriched with TF targets (TFT). TFT's are specific sets of genes that share a common TF-binding site defined in the TRANSFAC database [186]. TFT's are collected in the Molecular signature Database (MsigDB) [187] and are retrieved by WebGestalt upon analysis request. The examined promoter region has the size of -2kb, +2kb around the transcription start site. Then enrichment was evaluated through the hypergeometric test using the 10 most enriched terms with maximum significance level or p-value of 0.05. As we are testing multiple TFT families at the same time, the p values need to be adjusted for the effects of multiple testing, therefore we applied the sequential Bonferroni type procedure method proposed by [78]. We only considered enrichment of TFT families that were annotated for at least two genes. Finally, the results of the TFT enrichment analysis were visualized as heat maps to identify the common principles and differences of the enriched TF targets across the corresponding imprinted genes. This was done using the statistical language R [188].

4.3 Results

In this study we addressed the question whether imprinted genes as a group fulfill specific functions in mammalian organisms. For this, we tested if specific GO terms are overrepresented in the group of imprinted genes in comparison to all genes in the human or mouse genome. Of the 41 selected human imprinted genes, 38 are annotated in the GO database that contains in total 14116 human genes. In contrast, all 48 mouse imprinted genes are among the 14219 annotated mouse genes. One should note, though, that many genes are represented by more than one transcript in the GO database.

4.3.1 Imprinted genes are involved in developmental, transport and regulatory functions

First, we analyzed which terms of the Gene Ontology are enriched in the full set of all imprinted genes when compared to the set of all human genes or all mouse genes. We concentrate in this analysis on GO terms that are shared by at least 2 different imprinted genes. In this way, we assume to emphasize those cellular functions that relate to the controlled mono-allelic expression of the set of genes studied here.

In the human, the term *system development* is the term with the lowest p-value. This term is associated with 15 out of the 38 human imprinted genes. This corresponds to 2.6 fold enrichment in comparison to the annotation frequency in the group of all genes. *Cellular processes* is the term which is associated with the largest number of imprinted genes in the human: 32 imprinted genes (84.2% of all imprinted genes) are associated with this term, whereas this is only the case for 74.6% of all genes. For comparison, the imprinted genes in mouse showed a narrower range of 1.8 and 2 fold enrichment for these two broad terms, and only for *system development* the p-value is below 0.05. As shown in Table 4-1
Table 4-1, only the five generic GO terms, *multicellular organismal development, developmental process, neuron development, system development, and anatomical structure development* appear in both species with close to 2-fold enrichment (p<0.05, Fisher exact test). Only *neuron development* is 5-fold enriched.

As terms such as *system development* and *cellular processes* are rather general terms, we subsequently analyzed the enrichment of terms in the GO_FAT section of the DAVID database. As shown in Figure 4-1, among the enriched specific terms in human and mouse, some are linked to neuron development and differentiation and are intimately related with the CDKN1C and NDN genes. Interestingly, the terms *regulation of RNA metabolic process, regulation of transcription, DNA-dependent, and regulation of transcription* are the terms that are associated with the largest numbers of human imprinted genes (28.9, 28.9 and 34.2 %, respectively). Moreover, around 8.5% and 10.5% of the examined mouse imprinted genes are involved in the regulation process of phosphorylation and positive regulation of molecular function, respectively. This group includes the imprinted genes *Igf2, Ins2, Kcnq1, Htr2a, Grb10, Ndn, Tp73, Impact, Cdkn1c, Zim2, and Plagl1*.

The two GO terms *Regulation of RNA metabolic process* and the daughter node *Regulation of transcription, DNA-dependent* are associated with processes involved in the role of RNA synthesis regulation. Some of the encoded proteins are tumor proteins; others are inhibitors of the cell cycle, thus inhibiting division. It is also worth mentioning that the functional term *regulation of gene expression by genetic imprinting* (this is abbreviated to '*genetic imprinting'* in the DAVID database) is over-represented as well although it is associated in the Gene Ontology only with the genes *INS, IGF2*, and *KCNQ1* (Note: *INS* and *IGF2* are being interpreted by DAVID as a single locus which includes two alternatively spliced read-through transcript variants and align to the *INS* gene in the 5' region and to the *IGF2* gene in the 3' region).

Table 4-1 Conserved functional classes in imprinted genes in human (green) and mouse (brown) at adjusted p-value of 0.05.

Term	Species	Count	Percentage	Fold Enrichment	-Log (p-value)
GO:0007275 ~multicellular organismal development	Human	16	42.1	IIIIIIIIIIIIIIIIIIIIII 2.3	IIIIIIIIIIIIIIIIIIIIIIIIIIII 2.8
	Mouse	14	29.2	IIIIIIIIIIIIIIIIIII 1.9	IIIIIIIIIIIIIIIII 1.8
GO:0032502 ~developmental process	Human	17	44.7	IIIIIIIIIIIIIIIIIIIII 2.2	IIIIIIIIIIIIIIIIIIIIIIIIIIIII 2.9
	Mouse	15	31.3	IIIIIIIIIIIIIIIIIII 1.9	IIIIIIIIIIIIIIIIII 1.8
GO:0048666~neuron development	Human	4	10.5	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII 4.8	IIIIIIIIIIIII 1.3
	Mouse	4	8.3	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII 4.8	IIIIIIIIIIIII 1.3
GO:0048731 ~system development	Human	15	39.5	IIIIIIIIIIIIIIIIIIIIIIIIII 2.6	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII 3.3
	Mouse	12	25.0	IIIIIIIIIIIIIIIIIIII 2.1	IIIIIIIIIIIIIIII 1.7
GO:0048856 ~anatomical structure development	Human	15	39.5	IIIIIIIIIIIIIIIIIIIIIII 2.4	IIIIIIIIIIIIIIIIIIIIIIIIIIII 2.9
	Mouse	12	25.0	IIIIIIIIIIIIIIIIIII 1.9	IIIIIIIIIIIIII 1.5

These functional associations rely on publications about prominent imprinting control elements in the vicinity of these genes [189] and about epigenetic abnormalities in the *IGF2/H19* region of Beckwith-Wiedemann syndrome patients [190]. Furthermore, the GO term *genetic imprinting* that is a parent of the term *regulation of gene expression by genetic imprinting* is also annotated to the well-known imprinted genes *Gnas*, *NDN/Ndn* and *Peg3*. All in all, it is certainly fair to say that the coverage of genetically imprinted genes in the Gene Ontology is currently quite low.

Some functions related to transport are enriched and associated with both human and mouse imprinted genes. For instance, the Growth factor receptor-bound protein 10 (*GRB10*) is involved in the *Negative regulation of transport*. This gene interacts with insulin receptors and insulin-like growth-factor receptors [191]. Overexpression of some isoforms of *GRB10* inhibits tyrosine kinase activity and results in growth suppression, e.g. by suppressing glucose import [192].

The two enriched GO terms *Organic cation transport* and *Ion transport* describe the regulation of the directed movement of organic cations into, out of or within a cell, or between cells, by means of some agent such as a transporter or pore. The associated mouse imprinted genes *Slc22a2* and *Slc22a3* are polyspecific organic cation transporters in the liver, kidney, intestine, and other organs.

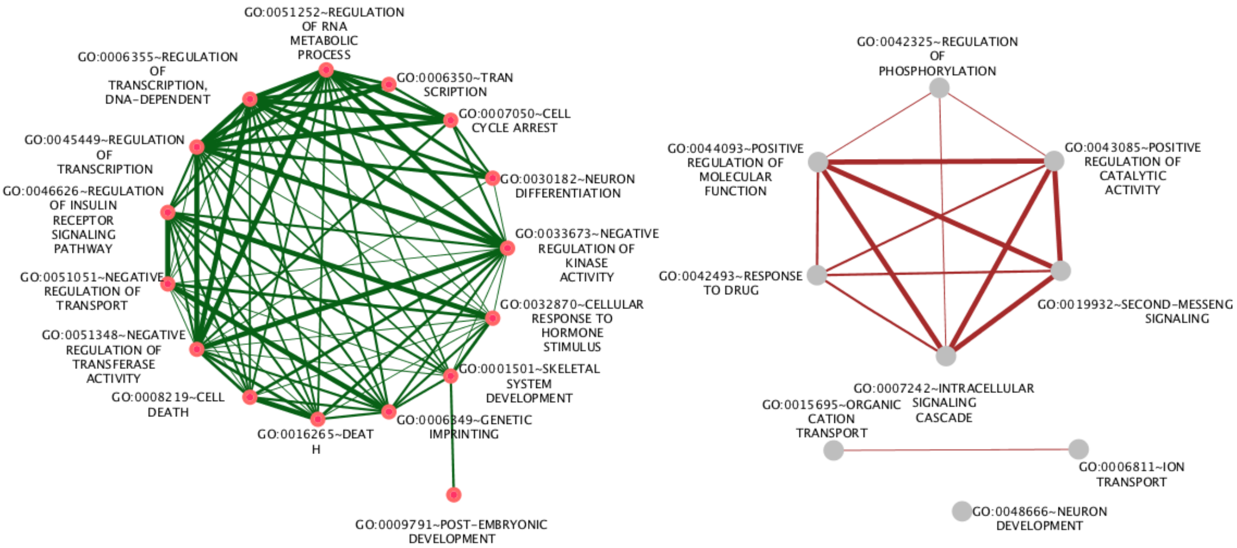

GO:0051252~REGULATION OF RNA METABOLIC PROCESS
GO:0006355~REGULATION OF TRANSCRIPTION, DNA-DEPENDENT
GO:0006350~TRAN SCRIPTION
GO:0007050~CELL CYCLE ARREST
GO:0045449~REGULATION OF TRANSCRIPTION
GO:0030182~NEURON DIFFERENTIATION
GO:0046626~REGULATION OF INSULIN RECEPTOR SIGNALING PATHWAY
GO:0033673~NEGATIVE REGULATION OF KINASE ACTIVITY
GO:0051051~NEGATIVE REGULATION OF TRANSPORT
GO:0032870~CELLULAR RESPONSE TO HORMONE STIMULUS
GO:0051348~NEGATIVE REGULATION OF TRANSFERASE ACTIVITY
GO:0001501~SKELETAL SYSTEM DEVELOPMENT
GO:0008219~CELL DEATH
GO:0006349~GENETIC IMPRINTING
GO:0016265~DEAT H
GO:0009791~POST-EMBRYONIC DEVELOPMENT
GO:0042325~REGULATION OF PHOSPHORYLATION
GO:0044093~POSITIVE REGULATION OF MOLECULAR FUNCTION
GO:0043085~POSITIVE REGULATION OF CATALYTIC ACTIVITY
GO:0042493~RESPONSE TO DRUG
GO:0019932~SECOND-MESSENG SIGNALING
GO:0007242~INTRACELLULAR SIGNALING CASCADE
GO:0015695~ORGANIC CATION TRANSPORT
GO:0006811~ION TRANSPORT
GO:0048666~NEURON DEVELOPMENT

Figure 4-1 The most specific enriched GO terms of biological functions for the full set of imprinted genes in human (green) and mouse (brown). Nodes represent the enriched Go terms and the thickness of the interconnected links corresponds to the number of shared genes.

Grouping genes based on shared GO terms can highlight functional similarities of different genes. For this, clustering algorithms were applied to a gene-to-gene similarity matrix and imprinted genes were classified into highly related groups (see methods). We identified one gene cluster in the human and two clusters in the mouse. The only discovered cluster in human resembles the second cluster in mouse and encompasses zinc finger protein genes such as *PEG3*, *ZNF597* and *ZNF331*. Its members have a strong association with regulatory and transcriptional tasks (Figure 4-2). For mouse, the first cluster contains mostly genes that encode proteins that are involved in transport processes (Figure 4-3a). As mentioned, the second group consists mostly of zinc finger protein genes similar to the human one (Figure 4-3b).

4.3.2 Maternally expressed genes dominate the role of imprinted genes in transport and gene regulation

In previous studies [184], Hutter et al. 2010 showed that maternally and paternally expressed genes differ in the level of conservation of their DNA sequences. For this reason, we analyzed whether maternally and paternally expressed genes differ also in their biological and molecular functions.

For the 19 maternally expressed genes in human, only 3 broad functional terms were found to be enriched, *nervous system development, organ morphogenesis*, and *positive regulation of osteoblast differentiation*. For the last GO term, the maternally expressed genes even showed a 59.4-fold enrichment although only two imprinted genes (*DLX5* and *GNAS*) are associated with this term. Thus, the enormous enrichment likely reflects that *positive regulation of osteoblast* is so far associated with very few genes in the full genome.

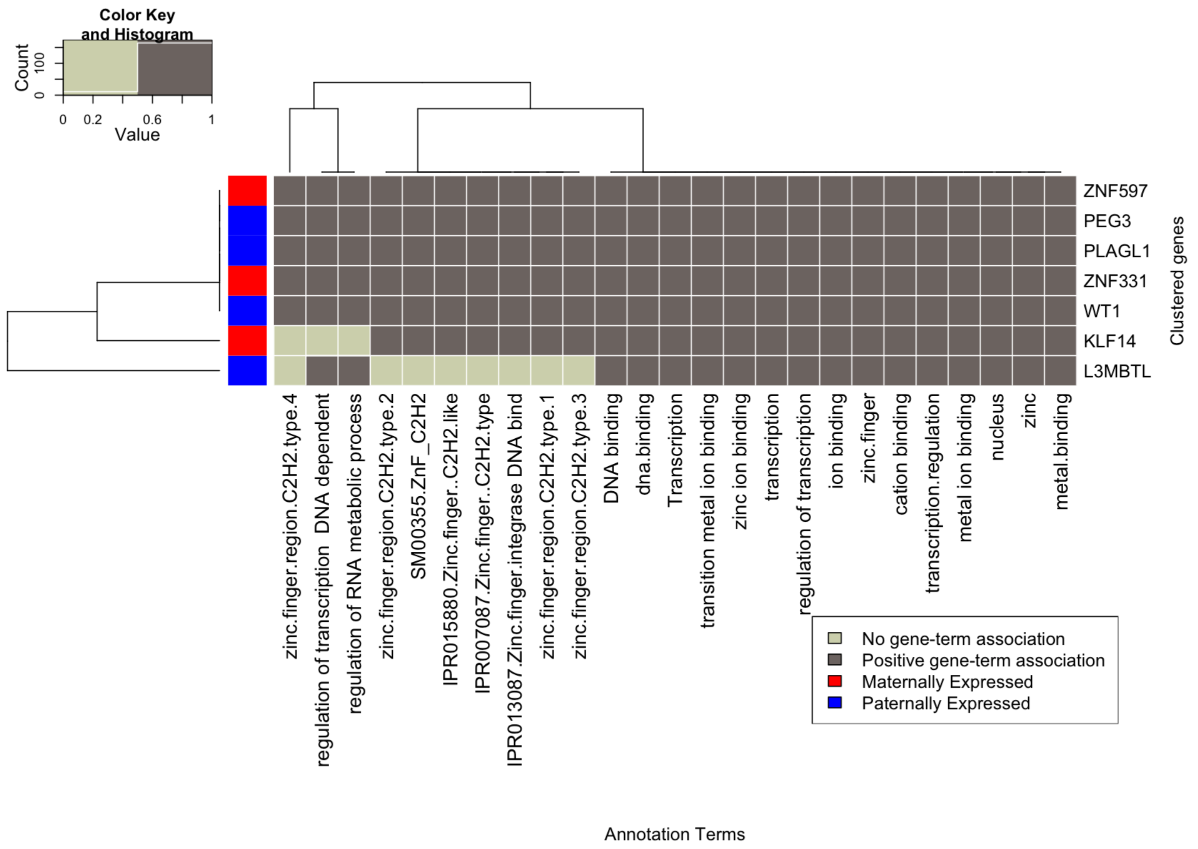


Figure 4-2 Functionally related imprinted genes in human.
The heat map view shows the gene-term association for those genes that share a high number of associated GO terms. Marked in red on the left side are maternally expressed genes; marked in blue are paternally expressed genes.

In mouse, 24 genes are classified as maternally expressed. We found that 14 biological functions are significantly associated with these genes. These 14 terms are dominated by a group of relatively unspecific terms related to transport processes such as *organic cation transport, transmembrane transport, ion transport* and *organic cation transport*. Therefore, not surprisingly, the five maternally expressed genes *Kcnk9, Kcnq1, Slca22a2, Slca22a3* and *Slca22a18* form a gene cluster that is associated with the same transport-related GO terms. The second gene cluster is formed by TF genes including the maternally expressed genes *Klf4* and *Zim1* (Figure 4-4).

4.3.3 Only few paternally expressed genes in human possess similar functions

The 17 paternally expressed genes in human are associated with fewer over-represented GO terms ($p<0.05$) than the maternally expressed genes. Most of them were already present in the over-represented terms for all imprinted genes (Figure 4-5). Thus we examined these genes on the basis of the GO_FAT knowledge base that contains more specific terms. Only two terms, i.e. *regulation of transcription, DNA-dependent* and *regulation of RNA metabolic process* are enriched for paternally expressed genes. Both terms are associated with the genes *PLAGL1, L3MBTL, IGF2, WT1, ZIM2,* and *PEG3*. Hence, both maternally and paternally expressed genes contain prominent groups of genes that have regulatory roles. Paternally expressed genes in mouse did not show any significant enrichment.

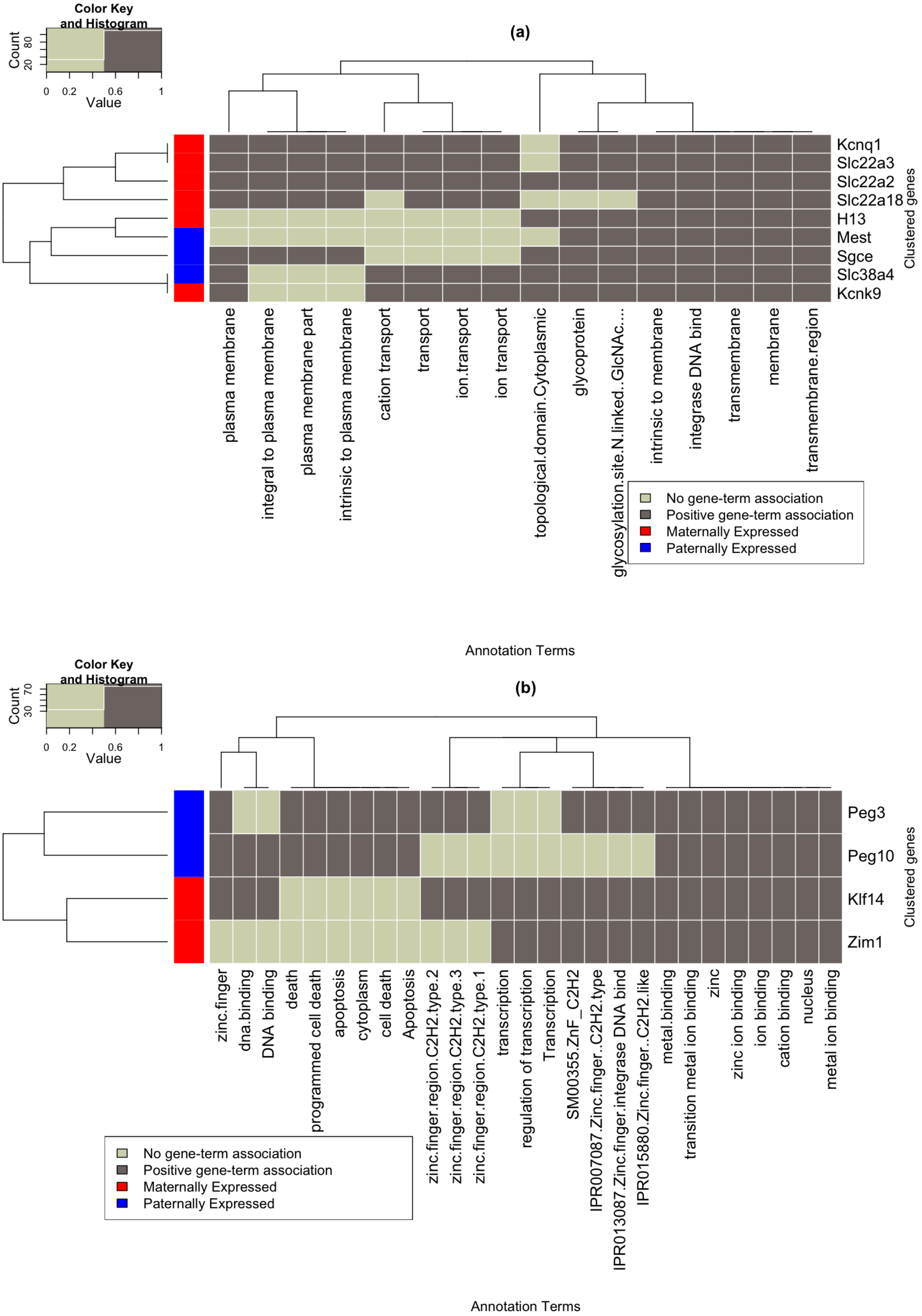


Figure 4-3 Functionally related imprinted genes in mouse.
Heat maps showing the gene-term association for the first and second gene clusters in Mouse. Marked in red on the left side are maternally expressed genes; marked in blue are paternally expressed genes.

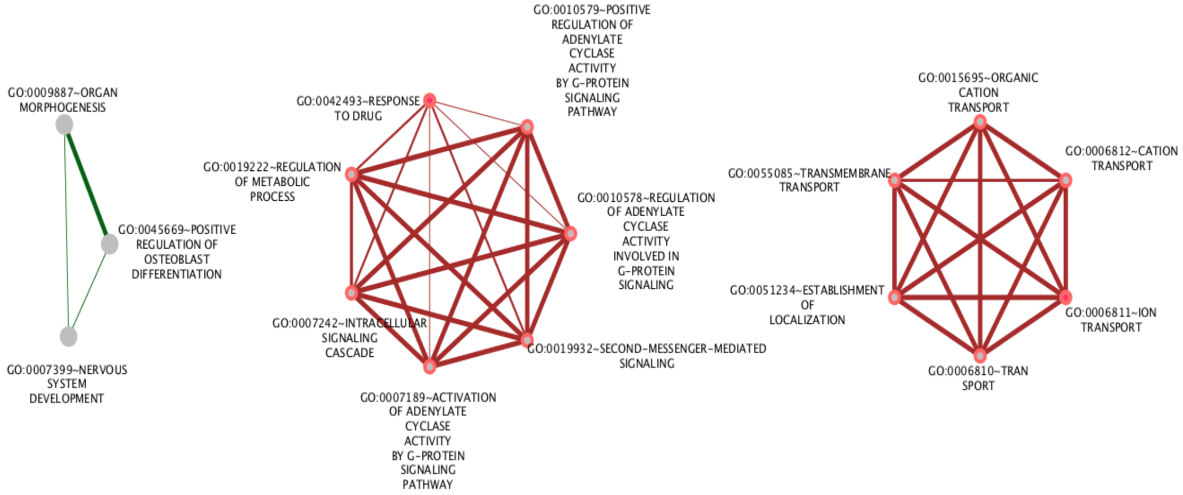

GO:0009887~ORGAN MORPHOGENESIS
GO:0045669~POSITIVE REGULATION OF OSTEOBLAST DIFFERENTIATION
GO:0007399~NERVOUS SYSTEM DEVELOPMENT
GO:0010579~POSITIVE REGULATION OF ADENYLATE CYCLASE ACTIVITY BY G-PROTEIN SIGNALING PATHWAY
GO:0042493~RESPONSE TO DRUG
GO:0019222~REGULATION OF METABOLIC PROCESS
GO:0010578~REGULATION OF ADENYLATE CYCLASE ACTIVITY INVOLVED IN G-PROTEIN SIGNALING
GO:0007242~INTRACELLULAR SIGNALING CASCADE
GO:0019932~SECOND-MESSENGER-MEDIATED SIGNALING
GO:0007189~ACTIVATION OF ADENYLATE CYCLASE ACTIVITY BY G-PROTEIN SIGNALING PATHWAY
GO:0015695~ORGANIC CATION TRANSPORT
GO:0006812~CATION TRANSPORT
GO:0055085~TRANSMEMBRANE TRANSPORT
GO:0051234~ESTABLISHMENT OF LOCALIZATION
GO:0006811~ION TRANSPORT
GO:0006810~TRAN SPORT

Figure 4-4 The enriched GO terms of biological functions for the maternally expressed genes in human (green) and mouse (brown). Nodes represent the enriched Go terms and the thickness of the interconnected links corresponds to the number of shared genes.

4.3.4 Enrichment analysis for the transcription factor targets

Mammalian genes are usually controlled by combinations of different TFs that bind to distinct binding sites in regulatory regions such as the promoters of genes. We were interested in the questions which TFs regulate imprinted genes and if paternally and maternally expressed genes can be distinguished by their TFs.

In total, we identified 25 TF families that showed an enrichment of binding sites in the set of imprinted genes in human ($p<0.01$, hyper-geometric test, see Methods). The associations between these families and the corresponding genes are shown in Figure A-1 (a) together with the expressed allele type. For mouse, binding sites for 40 TF families are enriched in imprinted genes at the same significance level of 0.01, see Figure A-1 (b). 19 transcription factor families possess binding sites that are enriched in the imprinted genes in both species (Figure 4-6). In species, *Nnat, Klf14, Blcap, Gnas*, and *Ube3a* are the genes that contribute most to the enrichment of transcription factor binding sites.

Figure 4-6 shows that in mouse and human, imprinted genes form similar, but not identical, clusters of genes that are regulated by the same transcription factor families. For example, the potassium channel genes *Kcnq1* and *Kcnk9* show an enrichment of heat shock factor 2 (HSF2) binding sites in human and mouse. Similarly, genes that are maternally expressed in placenta, such as *Slc22a18*, *Tfip2*, and *Phlda2*, cluster together in both species. In the mouse, this cluster is characterized by an enrichment of AP1 binding sites, whereas the prominent feature of the human gene cluster is a combination of AP1 and SP1 sites. Finally, Figure 6 illustrates clearly that paternally and maternally expressed genes do not cluster apart. This is also not the case if species-specifically enriched transcription factor binding sites are included (data not shown). Hence, paternally and maternally expressed genes are apparently not regulated by distinct combinations of TFs. and cannot be distinguished on a general level.

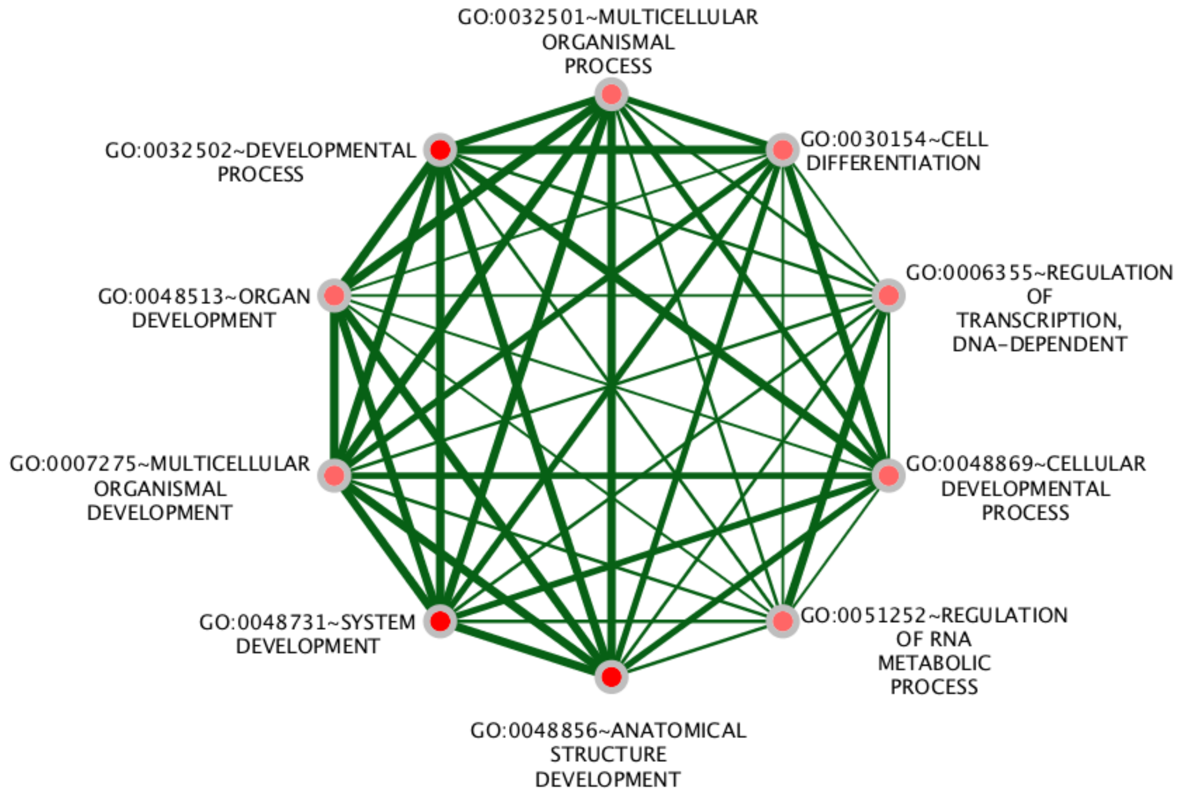

GO:0032501~MULTICELLULAR ORGANISMAL PROCESS
GO:0030154~CELL DIFFERENTIATION
GO:0032502~DEVELOPMENTAL PROCESS
GO:0006355~REGULATION OF TRANSCRIPTION, DNA-DEPENDENT
GO:0048513~ORGAN DEVELOPMENT
GO:0007275~MULTICELLULAR ORGANISMAL DEVELOPMENT
GO:0048869~CELLULAR DEVELOPMENTAL PROCESS
GO:0048731~SYSTEM DEVELOPMENT
GO:0051252~REGULATION OF RNA METABOLIC PROCESS
GO:0048856~ANATOMICAL STRUCTURE DEVELOPMENT

Figure 4-5 The enriched GO terms of biological functions for the paternally expressed genes in human. Nodes represent the enriched Go terms and the thickness of the interconnected links corresponds to the number of shared genes.

4.4 Discussion

This study analyzed enriched functional annotations of genetically imprinted genes based on the "biological process" tree of the Gene Ontology. In their seminal review [193], Tycko and Morrison concluded that the group of imprinted genes is predominantly involved in controlling growth and neurobehavioral traits. Tycko and Morrison pointed out that the numbers of paternally and maternally expressed genes related to growth are almost identical. On the other hand, only one maternally expressed gene (*UBE3A*) was linked to behavioral functions, in contrast to three paternally expressed genes (*SGCE, NDN, PWCR1*), as well as the paternally expressed genes *PEG1 (MEST)* and *PEG3* that were related both to growth and behavior. Thus, Tycko and Morrison argued that imprinting effects due to either maternally or paternally expressed genes are related to growth whereas behavioral functions are mostly controlled by paternally expressed genes. However, at the present stage, it is unclear if imprinted genes act indeed in the control of behavior, or if the observed behavioral abnormalities in mutant mice are caused by an impaired development of neurons and brain structures.

Our study did reveal an association of imprinted genes with developmental processes such as organ development in human and mouse. This indicates that these genes function indeed during embryogenesis, but they are not necessarily growth-regulating genes. The terms that are related to development in human as well as in mouse are associated with 25% to 44.7% of all imprinted genes in the respective species. Hence, a large proportion of imprinted genes contribute to developmental processes. Imprinted genes are also associated with GO terms that are related to neuronal development. Interestingly, neuronal development is apparently not a function that is restricted to paternally expressed genes. Furthermore, in comparison to developmental functions only a rather small number of imprinted genes (7 genes) show a functional association to the nervous system [194].

When paternally and maternally expressed genes are analyzed separately, mouse and human show clearly different associations. In the human, several maternally expressed genes (*DLX5, GNAS, TP73, PHLDA2, CDKN1C, PPP1R9A, UBE3A*) are associated with *organ morphogenesis*, and more particularly with *nervous system development* and *oesteoblast differentiation*. In the mouse, maternally expressed genes form two functional networks that are clearly separated. One is related to transport processes, and includes carrier proteins and channel proteins. Especially transport processes that are a key feature of placenta function are specifically associated with maternally expressed genes in the mouse. The second network consists of terms related to G protein signaling. This network is clearly dominated by *CALCR* and *SLC22A18*.

For the paternally expressed genes, a functional network is only found in the human. This network consists mostly of terms associated with development, and a few terms that are related to gene regulation. Interestingly, several imprinted genes that encode transcription factors (*PLAGL1, L3MBTL, WT1, ZIM2, PEG3*) seem to be key players in this network. Nevertheless, also among the maternally expressed genes are genes that regulate transcription. Thus, regulatory functions are not an exclusive feature of paternally expressed genes.

In this context we will briefly consider possible biases and shortcomings in the results obtained. While it is of course impossible to estimate how much we still don't know, even the annotations stored in the Gene Ontology clearly only represent a fraction of all knowledge in the original scientific literature. It is actually very difficult to provide an estimate how large this fraction is. As an example for this, only three out of 41 imprinted genes studied here are actually annotated in the GO as being "regulated by genetic imprinting" plus three that are related to "genetic imprinting". It is quite likely that the GO gives a more complete picture about the cellular functions of genes that have been studied intensely compared to the average gene. It is furthermore possible that some of the known imprinted genes such as *IGF2* belong to the group of intensely studied genes so that their cellular functions are known to a larger extent than those of less well studied genes and when compared to the average bi-allelically expressed gene. In agreement with this idea, we found that the three well-known genes *IGF2, INS,* and *GRB10* (out of 30) tended to dominate the functional enrichments in the group of paternally expressed genes. In contrast, the enrichments in the group of all imprinted genes were stable even when we removed the well-known genes *IGF2, INS*, and *GRB10*.

When grouping the imprinted genes by enriched GO annotations found for at least two genes, we applied the lowest recommended threshold value of 0.3. In future, when more complete functional associations will be available, it remains to be tested whether a higher, more cautious threshold would be advantageous.

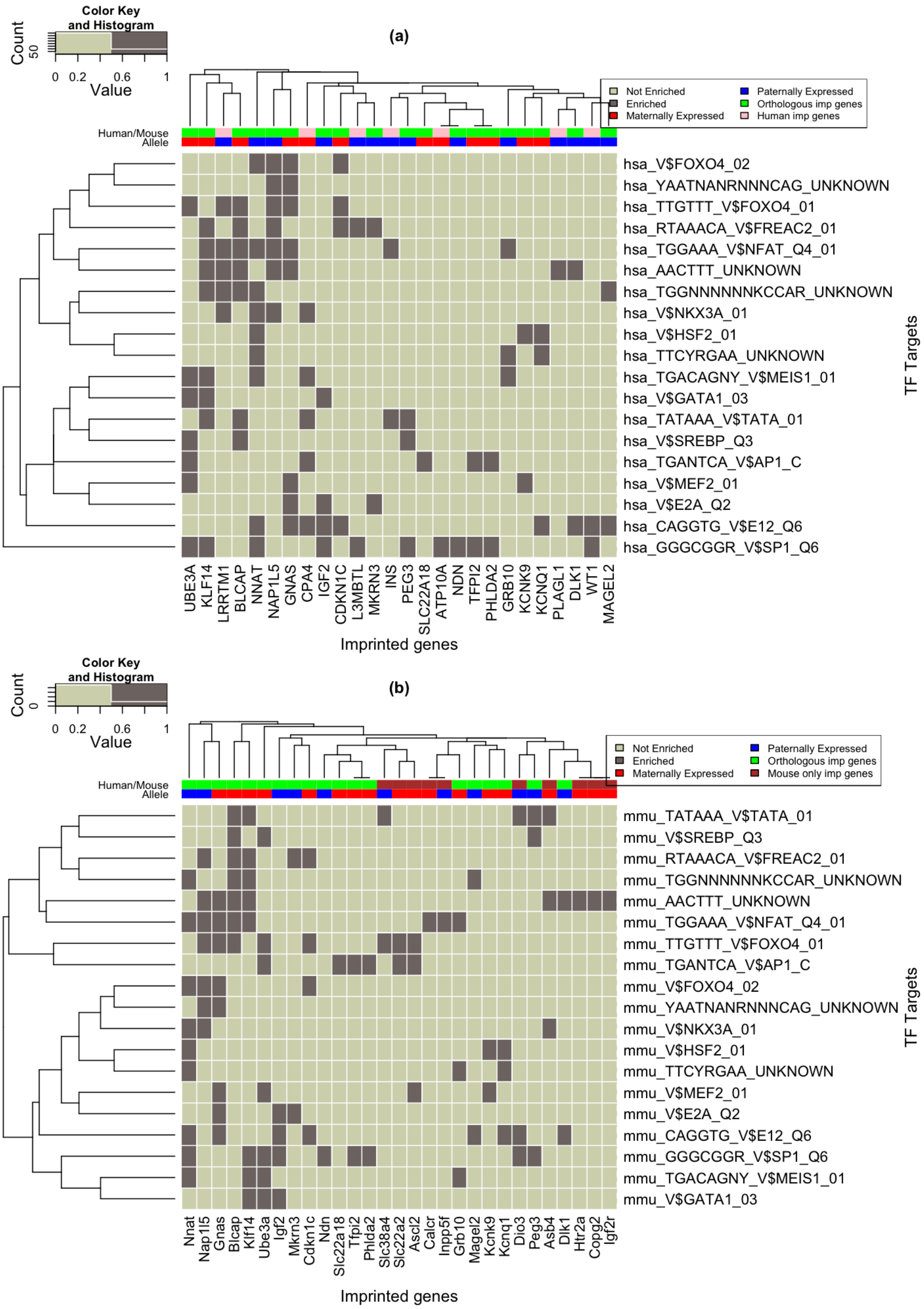


Figure 4-6 Conserved transcription factors in the full set of imprinted genes in human (a) and mouse (b) at adjusted p-value of 0.01. Marked in red and blue in the top line are the maternally, paternally expressed genes, respectively. Genes that are imprinted in both species are marked in green. Pink are the genes shown to be imprinted only in human, and brown are the genes shown to be imprinted only in mouse.

We found that when applied to the currently available data, this threshold gave a good compromise between coverage and specificity of the obtained results.

In the second part of the study, we were interested in the question if functionally related gene groups such as the prominent groups of transcription factors, and transport related proteins, are co-regulated by similar sets of transcription factor families. This is obviously not the case. Interestingly, also maternally and paternally expressed genes are not regulated by distinct sets of transcription factor families. In general, a few genes, i.e. *UBE3A, KLF14, BLCAP, NAP1L5, NNAT*, and *GNAS*, show an over-proportional enrichment of distinct transcription factor binding sites. Interestingly, these genes possess rather diverse functions. For example, *UBE3A* seems to act in neuronal development, whereas *GNAS* acts mostly in endocrinal pathways.

Although imprinted genes appear to be regulated by similar sets of transcription factors in mouse and human, it is difficult to identify a typical transcription factor that regulates imprinted genes. The most prominent factor appears to be SP1. This rather ubiquitous factor might be responsible for the broad tissue spectrum of imprinted genes [195]. On the other hand SP1 deficiency is to some extent associated with placental defects and impaired ossification, that are typical features of defects in imprinting [196].

Varrault and co-workers have recently identified a network of co-regulated imprinted genes involving the genes *Plagl1, Gtl2, H19, Mest, Dlk1, Peg3, Grb10, Igf2, Igf2r, Dcn, Gnas, Gatm, Ndn, Cdkn1c* and *Slc33a4* [197]. According to Fig. 6(b), E12 regulates four genes from this list (*Dlk1, Cdkn1c, Igf2* and *Gnas*); SP1 regulates three genes (*Peg3, Ndn* and *Igf2*) as well as AACTTT_UNKNOWN (*Igf2r, Dlk1* and *Gnas*). We suggest these three transcription factors as candidates that may be responsible for the coregulation of this imprinting network.

Berg and colleagues [198] recently analyzed the expression levels of ten of these genes (*Cdkn1c, Dlk1, Grb10, Gtl2, H19, Igf2, Mest, Ndn, Peg3*, and *Plagl1*) in mouse long-term repopulating hematopoietic stem cells and in representative differentiated lineages. Intriguingly, they found that most of the genes were severely down regulated in differentiated cells. They noticed that their study is the first one that connected imprinted genes that are known to be associated with embryonic and early postnatal growth to the regulation of somatic stem cells. Consequently, they suggested that the balancing forces of growth-promoting paternally expressed genes and of growth-limiting maternally expressed genes may as well play a role in keeping stem cells in the delicate balance of pluripotency. Along these lines, but in the opposite direction, our above finding that the global transcription factors E12 and SP1 play key roles in the regulation of imprinted genes fits to their well-known role in cell differentiation processes [199, 200].

5. Regulatory role of imprinted and pluripotency genes in hematopoiesis

This chapter is a shortened version of the following manuscript:

- Mohamed Hamed, Johannes Trum, Christian Spaniol, Mohammad R. Irhimeh, Martina Paulsen, and Volkhard Helms, Expression of pluripotency genes and imprinted genes during the onset of differentiation and during hematopoiesis [SUBMITTED].

Synopsis

The previous chapter discussed the functional roles of the imprinted genes and interestingly reported that many imprinted genes are transcriptionally regulated by hematopoiesis-related transcription factors such as NFAT, FOXO4, E2A, and TCF3. This has motivated the work presented in this chapter where we aimed at identifying regulatory elements from imprinted, pluripotency, and hematopoiesis associated genes that are putatively related to the transition of cells from the pluripotent stem-cell stage into the onset of development and into hematopoietic lineage commitment. To this end, we applied the GRN pipeline to gene expression data from three hematopoiesis–related datasets and one non-hematopoiesis–specific data set as a control.

Abstract

Maintenance of cell pluripotency, differentiation, and reprogramming are regulated by complex gene regulatory networks including monoallelically expressed, imprinted genes. Besides transcriptional control, epigenetic modifications such as DNA methylation and histone marks are increasingly gaining attention with respect to cellular differentiation. As a model system to study the onset of cell differentiation and subsequent cellular specialization, we have selected hematopoiesis and supplemented this with data from embryonic stem cell (ESC) lines. Using high throughput analysis of gene expression in mouse, the expression profiles of pluripotency-associated and imprinted genes were evaluated against known hematopoiesis-associated genes. We found that more than half of the pluripotency and imprinted genes are clearly upregulated in ESC and subsequently repressed. The remaining genes were either upregulated in progenitor or in differentiated cells. Thus, the three gene sets each consist of three similarly behaving gene groups with similar expression profiles in various lineages of the hematopoietic system as well as in ESCs. Co-expressed genes derived from the three gene sets were found to share gene ontology terms, which suggests a functional connection of the three sets during differentiation. To explain this coordinated behavior, we constructed a novel regulatory network of 32 imprinting-related genes that are shared with pluripotency or hematopoiesis genes. This network includes, among others, the genes *Myc, Nfkb1, Sp1, Sp3,* and *Tgfb1*, the regulatory gene *Oct4*, and *Wt1* and *Sp2* that regulate other genes that control pluripotency and hematopoiesis. This association suggests new aspects of the cellular regulation of the onset of cellular differentiation and during hematopoiesis involving, on the one hand, pluripotency-associated genes that were previously not discussed in the context of hematopoiesis and, on the other hand, involve genes that are related to genomic imprinting. These are new links between hematopoiesis and cellular differentiation and the important field of epigenetic modifications.

5.1 Introduction

The maintenance of cellular pluripotency, the onset of differentiation as well as cellular differentiation into particular lineages appear to be controlled by tightly regulated gene regulatory networks (GRNs) that describe the interactions between transcription factors (TCFs) and microRNAs and their target genes [201]. For mouse, Füllen, Schöler and co-workers have manually compiled from the original literature a dataset of genes termed the PluriNetwork that are involved in the regulation of the pluripotent state [202]. Besides transcriptional control, epigenetic modifications such as DNA methylation and histone marks are increasingly gaining attention with respect to cellular differentiation.

One of the hallmarks of epigenetics is the phenomenon of genomic imprinting, which describes parent-of-origin mono-allelic expression [179]. As the importance of epigenetic modes of gene regulation is particularly evident for imprinted genes, these genes serve as common model systems. Therefore, we are focusing here on the expression patterns and modes of regulation of the genes that have been shown to be mono-allelically expressed in the mouse.

Hematopoiesis describes the differentiation of hematopoietic stem cells (HSCs) into lineage-committed progenitor cells. Recent transcriptomics studies have identified important parts of the regulatory networks that control maintenance of HSCs [203] and progenitors [201, 204, 205]. Despite the fact that HSCs share the hallmark properties of long-term self-renewal and multi-lineage differentiation capacity, it has been shown that their chromatin state and the expression patterns of TCFs do vary substantially based on the location of HSCs in bone marrow, the origin (i.e embryo, adult, or aged) and time of study [206]. Still, some parts of the GRN architecture are expected to be conserved in the different hematopoietic lineages [206].

Not much is known about the imprinting status of imprinted genes during blood cell differentiation. As an exception to this, maternal imprinting at the H19-Igf2 locus was shown to maintain adult haemotopoietic stem cell quiescence [207]. Besides, several lines of evidence do exist for the importance of imprinted genes during the transition from the stem cell stage to differential commitment as well as during particular cell lineages, namely hematopoiesis. For example, a network of 15 co-regulated imprinted genes involved in embryonic growth has been identified [197]. 10 of these genes were down regulated in terminally differentiated mouse cells compared to long-term repopulating HSCs [198]. In multipotent progenitor cells, 13 out of 15 imprinted genes were clearly downregulated compared to HSC whereas the two imprinted genes *Gnas* and *Gatm* were upregulated in MPP3 and MPP4 relative to MPP1 and HSC [204]. Recently, we have identified 10 imprinted genes that are transcriptionally regulated by the hematopoiesis related TCF NFAT. We also found 9 imprinted genes that are targets of *FOXO4* TCF.[208] In CD34$^+$ cells, the imprinted maternally expressed gene *p57* (*Cdkn1c)* was the only cyclin-dependent kinase inhibitor to be rapidly up-regulated by TGFβ, a negative regulator of hematopoiesis [209]. Additionally, we found that promoter regions around the transcription start sites of *Mkrn3*, *Igf2*, and *Gnas* genes contain DNA motifs that match to annotated binding site motifs for the TCFs *E2A* and *TCF3*. The latter plays major roles in determining tissue-specific cell fate during embryogenesis such as early B lymphopoiesis and germinal center B-cell development [210]. Several studies from the Li group indicated that the expression of certain imprinted genes changes in HSCs during differentiation from quiescent to multi-lineage progenitors [211]. However, the transcriptional activity of imprinted genes and imprinting-related genes that are regulators of imprinted genes in the onset and further progression of cell differentiation (on the example of hematopoiesis) and the aspects of their involvement have not been addressed in detail before.

In this study, we aim at identifying regulatory elements that are putatively related to the transition of cells from the pluripotent stem-cell stage into the onset of development and into lineage commitment. In order to characterize the involvement of imprinted genes and pluripotency-associated genes during murine hematopoiesis in a systematic way we have re-analyzed previously deposited microarray datasets from different stages of hematopoiesis and from embryonic stem cells (ESC). The expression patterns of imprinted genes and pluripotency-associated genes during these stages were compared to the global expression patterns of hematopoiesis-associated genes and we set out to explain how the similarity of the gene expression arises. Our results suggest that imprinted genes, that are known to be associated with embryonic and early postnatal growth, may as well play a collaborative role (i) in keeping stem cells in the delicate balance of pluripotency and (ii) during the onset of cell differentiation towards hematopoiesis.

5.2 Methods

5.2.1 Genes selection

Three mouse gene lists were prepared (imprinted, pluripotency, and hematopoiesis). Our selection of imprinted genes was not done manually as in chapter 4 and in the thesis work of Barbara hutter [208, 212, 213] as the manually curated lists contained a rather restricted number of genes. In this study, our selection was based on the overlap of several well-known online catalogs of imprinted genes. Imprinted genes were downloaded in July 2012 from four well-known catalogs [IGC database (http://www.otago.ac.nz/IGC) [185], Geneimprint (http://www.geneimprint.com/site/genes-by-species.Musmusculus), WAMIDEX (https://atlas.genetics.kcl.ac.uk),[214] and MouseBook™ (http://www.mousebook.org/catalog.php?catalog=imprinting)]. Then a single list of 120 genes (called henceforth candidate imprinted genes) was created from the four catalogs by including only genes that appeared in at least two catalogs and by filtering out genes that have conflicting or unknown imprinting status in the various catalogs (i.e whether they are maternally or paternally expressed). 86 imprinted genes were present on the microarray chip. As this is a computational study, we did not verify experimentally whether these genes are actually mono-allelically expressed.

The pluripotency list including 274 genes was obtained from the PluriNetWork [202], a hand curated pluripotency-controlling gene network in mouse with 574 regulatory interactions. To the best of our knowledge, no generally accepted GRN for the global hematopoiesis system has been established. In the absence of such a model, we considered as hematopoiesis genes the 615 genes that are annotated in the Gene Ontology [67] for the GO term *hematopoietic or lymphoid organ development* (GO:0048534). Not all genes in the three gene lists were annotated in the Affymetrix array. Of the 120 imprinted genes only 86 were annotated (the rest were mostly non-coding RNAs, which are thus not considered), whereas only 2 out of 274 pluripotency genes and 53 out of the 615 hematopoietic genes were not annotated. The counts of overlapping genes are shown in Figure 5-1.

5.2.2 Microarray analysis

Gene expression microarray data [three hematopoietic datasets (accession IDs GSE6506 [215], GSE14833 [216], GSE34723 [217]) and one non-hematopoietic specific (control) (GSE10246 [218]) that also contains ESC samples] generated with Affymetrix GeneChip Mouse Genome 430 2.0 Array were downloaded from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) [219].

Data normalization, model-based expression measurements, and annotation of the imprinted and pluripotency genes to their corresponding probes in the four datasets were done using the GC-RMA method and mouse 4302.db packages, respectively, by using the Bio-conductor software [220] within the statistical programming language R [188].

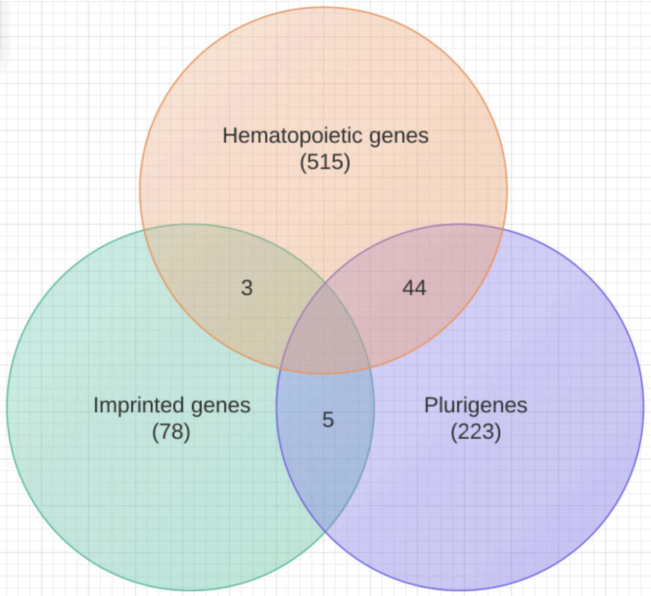


Figure 5-1 Venn diagram of the 3 gene sets involved in the analysis. Imprinted, pluripotency, and hematopoietic genes.

A gene expression similarity score was calculated to test how similar the normalized expression of an individual gene (in the chip including imprinted genes) is to the distribution of normalized expression values for the sets of pluripotency and hematopoiesis genes separately across the four datasets, see supplementary material. When considering the similarity to pluripotency genes, for example, the expression value of a gene g_i in a cell sample s_j was weighted by the number of pluripotency genes having the same expression value in the same sample PD_{sj}. This product was summed over all samples to give a representative score for each gene.

$$SimScore\ (gi) = \sum_{j=1}^{cell\ samples} PD_{sj}\left(gi_{sj}\right) \qquad where\ i\ \in [1, imprinted\ genes\ count]$$

$$and\ j\ \in [1, all\ cell\ samples\ per\ dataset]$$

Next, we separated the similarity scores of imprinted genes and non-imprinted genes and examined with the Mann-Whitney U-test whether imprinted genes have a higher gene expression similarity to pluripotency and hematopoiesis genes than the background of all other genes. Additionally, we defined top scored genes as the highest 10% of the ranked genes and then applied the hyper-geometric test to investigate the significance of having imprinted genes among the defined top scored genes.

For lineage specificity, six isometric lineages (three lymphoid and three myeloid) were constructed from the four expression datasets by following the hematopoietic differentiation model in [217] (We looked at three main hematopoiesis developmental stages: early progenitors (LTHSC and STHSC), intermediate progenitors (LMPP and CLP), and terminally differentiated blood cells (MKE and GM). Then we used a conservative differential expression approach based on moderated t-test to encode the differences between the three stages.

Table 5-1). We looked at three main hematopoiesis developmental stages: early progenitors (LTHSC and STHSC), intermediate progenitors (LMPP and CLP), and terminally differentiated blood cells (MKE and GM). Then we used a conservative differential expression approach based on moderated t-test to encode the differences between the three stages.

Table 5-1 Selected hematopoietic lineages and their stages of sequential cell development.
HSC, Hematopoietic Stem Cells; *MPPa*, Multipotent Progenitor state A; *MPPb*, Multipotent Progenitor state B; *GMLPa*, Granulocyte Macrophage Lymphoid Progenitor state A; *GMLPb*, Granulocyte-Macrophage-Lymphoid Progenitor state B; *CLP*, Common Lymphoid Progenitor; BLP, Earliest B Lymphoid Progenitor; *PREPROB*, Precursor of B-cells Progenitor; FrB, Fraction B B-cell; FrC, Fraction C B-cell; FrD, Fraction D B-cell; FrE, Fraction E B-cell; iNK, intermediate Natural Killer Cell; mNK, mature Natural Killer Cell; DN1, Double Negative T-cell 1; DN2, Double Negative T-cell 2; DN3a, Double Negative T-cell 3a; DN3b, Double Negative T-cell 3b; DN4, Double Negative T-cell 4; DPCD69-, Double Positive CD69- T-cell; DPCD69+, Double Positive CD69+ T-cell; CD4+CD69+, CD4+ CD69+ T-cell; CD4+CD69-, CD4+ CD69- T-cell; MEP, Megakaryocyte/ Erythrocyte Progenitor; pMEP, pre of MEP; pCFU-E, pre of CFU-E; sCMP, Strict Common Myeloid Progenitor; GMP, Granulocyte-Macrophage-Progenitor; pGMPa, pre of GMP state A; pGMPb, pre of GMP state B; Mono, Monocytes; Gra, Granulocytes.

Lineage	**Sequential Cell Development**
B-cell	*HSC→ MPPa → MPPb→ GMLPa → GMLPb → CLP→ BLP→ PREPROB → FrB→ FrC → FrD → FrE*
NK-cell	*HSC→ MPPa→ MPPb→ GMLPa→ GMLPb→ CLP→ iNK→ mNK*
T-cell	*HSC→ MPPa→ MPPb→ GMLPa→ GMLPb→ CLP→ DN1→ DN2→ DN3a→ DN3b→ DN4→* $DPCD69^{-}$ → $DPCD69^{+}$ → $CD4^{+}CD69^{+}$ → $CD4^{+}CD69^{-}$
Erythrocytes	*HSC→ MPPa → pMEP → MEP → pCFU–E*
Monocytes	*HSC → MPPa → sCMP→ pGMPa→ pGMPb → GMP → Mono*
Granulocytes	*HSC→ MPPa → sCMP → pGMPa → pGMPb → GMP → Gra*

P-values were adjusted using Benjamin- Hochberg procedure [78] to limit the false discovery rate to 5%. In order to alleviate the typical loss of statistical power resulting from performing multiple testing on a gene-by-gene basis, we performed non-specific pre-filtering by selecting genes on the basis of variability before the differential analysis. We removed 20% of all genes showing the least variability across lineages in the datasets and kept only genes that showed higher variation across the lineage and are thus potentially good candidates for differentially expressed genes [221].

Lineages that are constructed from GSE6506 dataset and contain only two stages (early progenitors and terminally differentiated cells) were analyzed by setting on/off expression threshold (similar to [215]) to identify uniquely expressed genes in each stage of the cell development of each lineage. Finally, a gene was confirmed as differentially expressed gene if it appeared in the same lineage in at least two different datasets.

5.2.3 Reconstruction of an imprinted gene network (IGN)

Gene expression data of the four accession IDs were subjected to weighted gene co-expression network analysis for describing the correlation patterns among genes across the 67 considered microarray biological samples. The popular hierarchical clustering (HCL) method was used for clustering taking Pearson correlation as a distance metric.

The WGCNA package [81] was employed to map the strength of gene pair interconnections to proportional edge weights and to produce a module centric co-expression network. This co-expression network of imprinted genes was subsequently expanded by (a) including additional genes coding for TCFs that regulate any of the considered imprinted genes, and (b) by including target genes that are regulated by any of the imprinted genes acting as TCFs themselves and then called "imprinted gene network" (IGN).

5.2.4 Functional enrichment and similarity

The functional annotation tool in DAVID was used to identify significantly enriched functional categories in gene sets [135]. We determined which GO categories that were annotated to at least 2 genes and are statistically overrepresented in the co-expressed genes against the full mouse genome (control). Enrichment was evaluated through the Fisher Exact test using *p*-value threshold of 0.05. Functional similarity between each pair of genes was measured by FunSimMat [222] (http://funsimmat.bioinf.mpi-inf.mpg.de/help8.php) and GO terms were visualized as a scatter plot by REVIGO [223].

5.3 Results

In this study, we re-analyzed published gene expression microarray data deposited in GEO [219] [three hematopoietic datasets (accession IDs GSE6506 [215], GSE14833 [216], GSE34723 [217]) and one non-hematopoietic specific (control) (GSE10246 [218])]. As explained in the methods section, we established three gene lists of imprinted, pluripotency-associated, and hematopoiesis-associated genes. In the remainder of this chapter, we will use the short names "pluripotency genes" and "hematopoiesis genes" while noting that, e.g. some genes in the pluripotency list might be directly involved in maintaining the pluripotency of ES/iPS cells, whereas some genes might have indirect and more general functionalities, such as cell cycle regulators etc. From these lists, 86 imprinted, 272 pluripotency and 562 hematopoietic genes are annotated on the microarray.

5.3.1 Imprinted genes show similar expression patterns to pluripotency and hematopoiesis genes

To get an overview, Figure 5-2 shows clustered normalized expression profiles for two ESC lines, three progenitor cell lines (Long Term HSC: LTHSC, Common Myeloid Progenitor: CMP, and Granulocyte-Macrophage-Progenitor: GMP), and three terminally differentiated cell lines (Nk-cells, B-cells, T-cells). Clustering as well as visual inspection revealed three main classes of expression patterns, which are shared by most imprinted, pluripotency and hematopoietic genes. The first class contains genes that have high expression levels in ESC and have gradually decreasing expression levels during the two stages of hematopoiesis (early and intermediate progenitors and terminally differentiated blood cells). As expected, more than half of the imprinted genes (left panel, green) and of the pluripotency genes (middle panel, blue) belong to this class. Also, about one third of the hematopoiesis genes (right panel, orange) belong to this class. Genes of the second class are characterized by over-expression in the early and intermediate progenitors (more specifically in Common Lymphoid Progenitor*: CLP*) and relatively lower levels in both ESC and terminally differentiated cells. The third class includes genes that are predominantly upregulated in matured blood cells.

Interestingly, the second and third classes contain genes from all three-gene sets. On the basis of gene ontology (GO) annotation, we investigated the functional similarity of the three genes sets among each other and with respect to randomly selected genes. This analysis revealed that pluripotency genes and hematopoiesis genes share the highest similarity of GO annotation. This is quite expected since the genes from both sets are involved in regulating cell fate. No difference was found when the functional similarity of pluripotency genes belonging to class 1 was compared to hematopoiesis genes also belonging to class 1, or when comparing the similarity between mixed classes. In comparison, the average functional similarity of imprinted genes with pluripotency genes or with hematopoiesis genes was lower (about 0.6), but still clearly higher than that with randomly selected genes.

To put this visual impression onto a quantitative basis, we then ranked all genes according to their gene expression similarity score across all considered hematopoietic samples. Notably, all p-values for the three hematopoietic datasets (that encompass differentiation and cell development data only) were significant (between 0.001 and 0.01). Moreover, a large portion of imprinted genes belongs to the highest 10% of the ranked genes in GSE6506 and GSE34723 datasets (66 % and 59 % respectively). In contrast, no significant difference was found between the ranking of imprinted genes and the background of all genes of the control (GSE10246; largely non-hematopoietic) and the number of top ranked imprinted genes was lowest here.

Table 5-2 Genes' similarity scores statistical comparison.
Mann-Whitney U-test was used to test if imprinted genes have a higher gene expression similarity to pluripotency and hematopoiesis genes than the background of all other genes (non-imprinted genes). Then genes the ranked top 10% scoring genes were tested using hyper-geometric test to find out the significance of having imprinted genes among the defined top scored genes. (*) Among the three consistent datasets, only the p-value of hyper-geometric test for GSE14833 does not meet the significance threshold of 0.05.

Dataset	Compared genes to background	Mann-Whitney U-Test	Top Scored Imprinted Genes	Hyper-geometric Test
GSE6506[215]	**Pluripotency**	0.006	55	0.006
	Hematopoietic	0.044	57	0.010
GSE34723[217]	**Pluripotency**	0.004	50	0.004
	Hematopoietic	0.003	51	0.009
GSE14833[216]	**Pluripotency**	0.003	18	0.195 *
	Hematopoietic	0.006	24	0.214 *
GSE10246[218]	**Pluripotency**	0.106	11	0.784
(Control)	**Hematopoietic**	0.101	14	0.700

5.3.2 All three gene sets contribute to hematopoietic lineage specificity

To discover if there are proteins encoded by imprinted, pluripotency, and hematopoietic genes that act during differentiation of particular lineages, we subjected the selected microarray datasets to differential analysis. We studied genes that change their expression patterns during the sequential stages of cell development of specific lineages (Table 5-1)

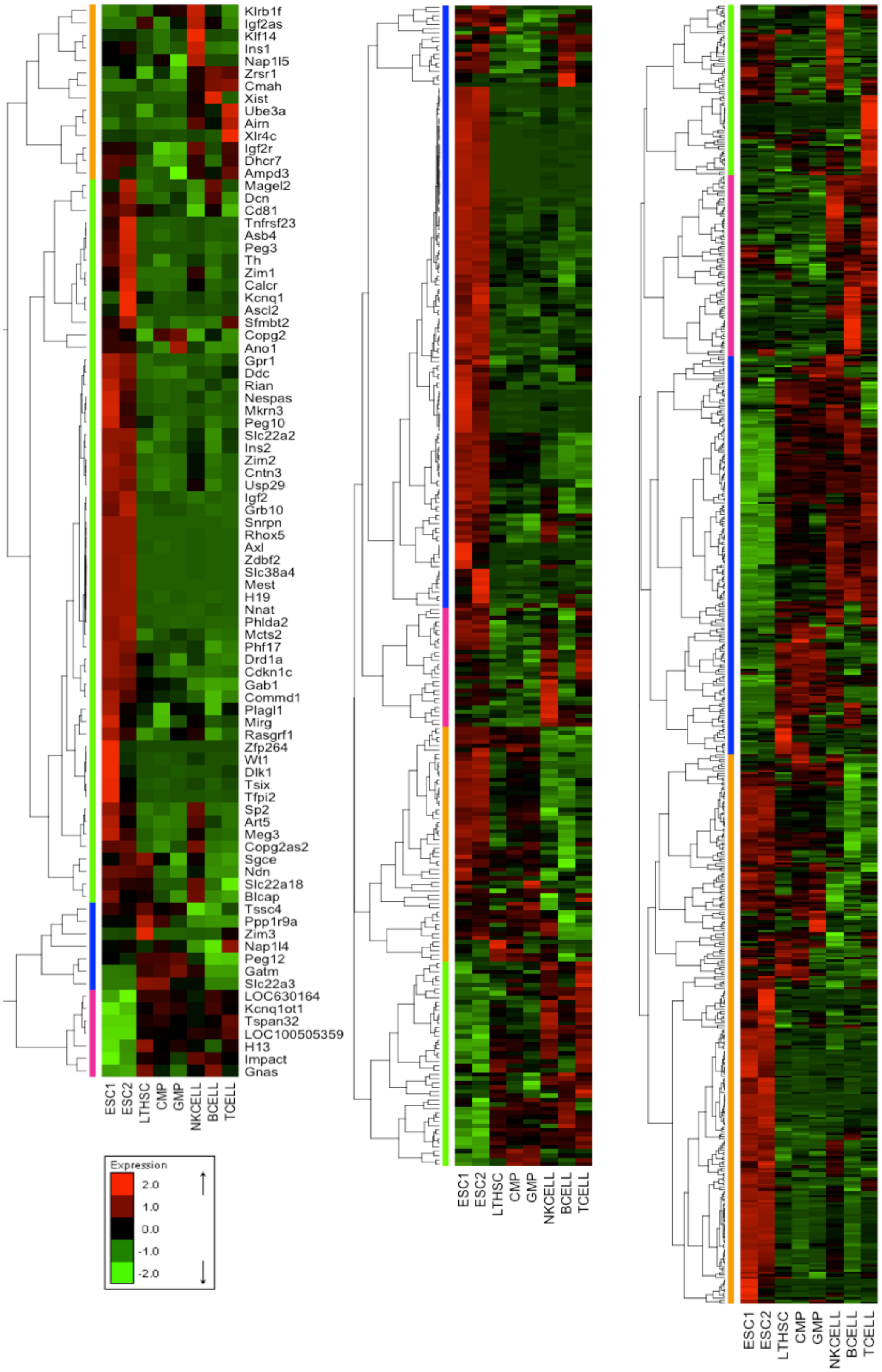


Figure 5-2 Heatmaps showing transient changes in expression profiles.

Different groups of ESC and hematopoietic cells (e.g stem cells, intermediate progenitors, and terminally differentiated blood cells) from the GSE10246 dataset for (left panel) imprinted genes, (middle panel) pluripotency genes and (right panel) hematopoietic genes were compared. Green spots represent down-regulated genes, and red spots represent up-regulated genes. The order of genes is obtained by hierarchical clustering, which shows three similar pattern classes between imprinted, pluripotency and hematopoietic genes.

For the purpose of differential expression analysis, we divided cell samples into three classes [early progenitors (e.g *HSC* and *MPP*), intermediate progenitors (e.g *GMLP* and *CLP)*, and terminally differentiated blood cells (e.g *Monocytes* and *Nk-cells*)]. This analysis was now based on far more cell types than the global analysis of Figure 5-2.

Lineage-specific differentially expressed genes (here termed marker genes) found in the three gene sets and the related expression heatmaps for NK-cells, monocytes, and erythrocytes are shown in Figure 5-3 (heatmaps for the other 3 lineages are shown in supplementary Figure B-1). The number of significant lineage markers varies between 23 genes (in granulocytes) and 193 genes (in B-cells). Only the three genes *Rbp1* and the two imprinted genes *Sgce* and *Mkrn3* are shared by all myeloid branches (erythrocytes, monocytes, and granulocytes) (Supplementary Table B-1). Additionally, we identified 16 marker genes (e.g *Lgals1, Gimap5, Pml,* and *Hoxa5*) that are exclusively differentially expressed in myeloid lineages (not in lymphoid). These 16 genes are annotated for terms like GO:0002317 "*plasma cell differentiation*", GO:0043011" *myeloid dendritic cell differentiation*", GO:0030099 " *myeloid cell differentiation*", and GO:0045639 "*positive regulation of myeloid cell differentiation*", respectively. Along the same lines, the lymphoid markers contain 30 genes shared by all lymphoid peers (B-cell, T-cell, and NK-cell) and 226 genes that were only detected for individual lymphoid lineages (not myeloid) such as *Tcf7*, *Lef1*, and *Rel*, which plays a role in differentiation and lymphopoiesis [224]. Remarkably, most differentially expressed genes in B-cells (102 genes) and in T-cells (70 genes) belong to the pluripotency genes (Supplementary Table B-1) and a large portion of them was imprinted (27, and 30, respectively), whereas the large hematopoiesis set (516 genes) contributes only 64 and 53 differentially expressed genes, respectively.

Separate labeling of maternally and paternally expressed genes did not reveal a clear-cut separation, which is consistent with previous findings [208]. Nevertheless, only paternally expressed genes were differentially expressed in the erythrocyte lineage (Figure 5-3). In contrast, the imprinted genes that are overexpressed during late stages of hematopoiesis tend to be maternally expressed (e.g *Cmah* and *Nap1l4* in B- and T-cells, *Klrb1f* in monocytes and NK-cells, *Th* and *Igf2r* in T-cells) rather than paternally expressed (*Sp2, Mcts2,* and *Ddc* only in B-cells and T-cells). Three imprinted genes (Ndn, Peg3, and Peg12) that were annotated by Chambers and colleagues [215] as HSC specific genes were identified here as marker genes for differentiated lineages. Consistent with the findings of Chambers et al., they are highly expressed in HSCs and downregulated in differentiated states.

The postulated functional role of the identified lineage markers during hematopoiesis was backed up by inspecting the mammalian phenotypes associated with hematopoiesis abnormalities using the MGI database [225], Supplementary Table B-1. Apparently, lineage-specific genes show deficiencies in either functionalities or differentiation of a specific lineage, validating the used approach in identifying the lineage markers. An example from the B-cell lineage is the knockout of the imprinted gene *CD81.* This is reported to cause *abnormal B cell morphology* (MGI ID: MP:0004939), *decreased B-1 B cell number* (MP:0004978), and *instability in B cell proliferation* (MP:0005154, MP:0005093). More generally, the knockout of the imprinted gene *Cdkn1c* leads to *decreasing hematopoietic stem cell number* (MP:0004810) and *abnormal hematopoietic stem cell physiology* (MP:0010763). From the set of pluripotency genes, gene knockout of *Relb* exhibits also several abnormalities such as *decreased B cell number* (MP:0005017), *decreased B cell proliferation* (MP:0005093), *absent lymph nodes* (MP:0008024), *decreased pre-B cell number* (MP:0008209), and *extra-medullary hematopoiesis* (MP:0000240).

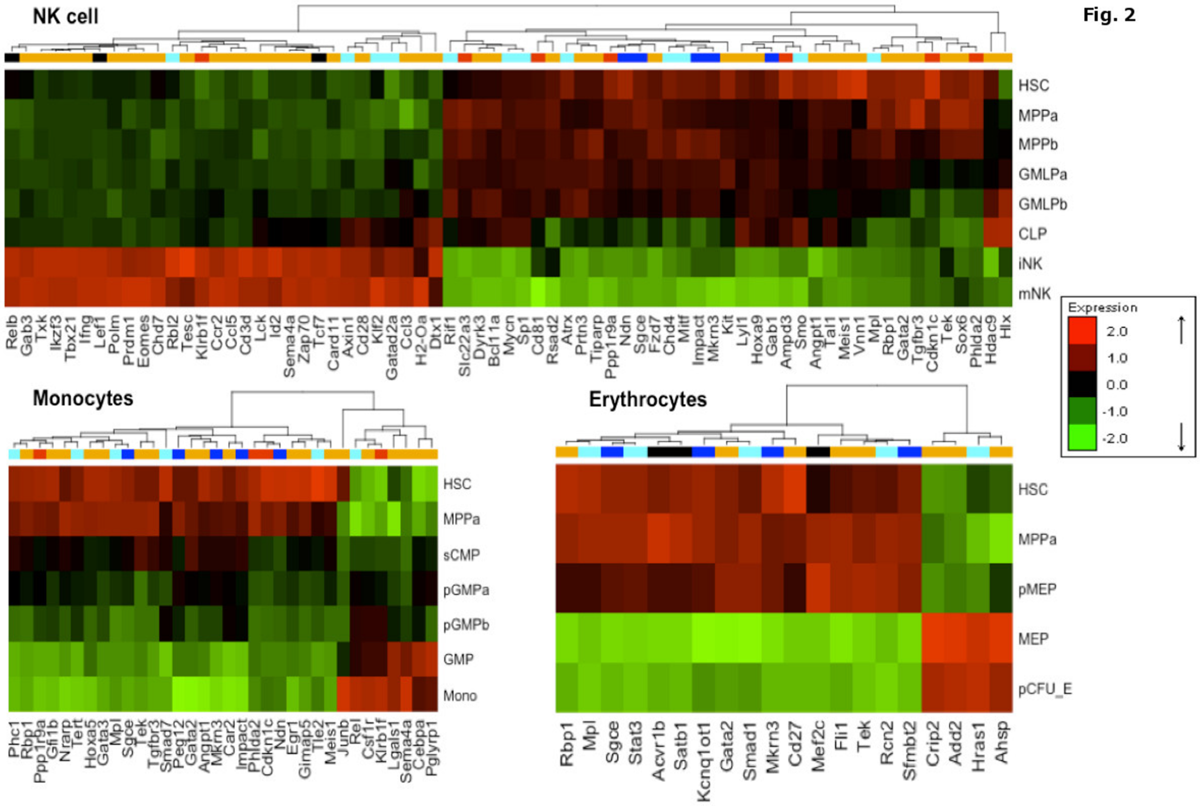


Figure 5-3 Heatmaps of differentially expressed imprinted genes.
The order of genes is obtained by hierarchical clustering of three blood lineages (NK-cells, Monocytes, and Erythrocytes) based on the GSE34723 dataset. Gene clustering color coding is (blue) for paternally expressed genes, (red) for maternally expressed, (cyan) for pluripotency genes, and (orange) for hematopoietic genes. The other three lineages (B-cells, T-cells, and granulocytes) are shown in the supplementary Figure B-1. Shared genes between the pluripotency and hematopoietic gene sets are marked in black. Green spots represent down-regulated genes, and red spots represent up-regulated genes. The clustering reveals that for every lineage, there exist imprinted as well pluripotency and hematopoietic genes showing similar expression changes during cell development.

Finally, we analyzed the functional similarity of the identified imprinted genes either to pluripotency or hematopoietic genes using the tool FunSimMat [222]. This was done separately for each lineage in comparison to the similarity values of the background genes that are not differentially expressed in the corresponding lineage. Interestingly, we found that lineage specific genes from the gene set pairs (imprinted-pluripotency and imprinted-hematopoietic) have an elevated functional similarity between 0.4 and 0.6 to each other for the biological process (BP) category in comparison to that between the other genes in the two gene set pairs (Supplementary Figure B-2a). The functional similarity scores between imprinted and hematopoietic genes were ~0.35 to 0.75 (p-values 0.178 to 6.0E-237) (only in erythrocytes lineage, marker genes did not show a significantly different functional similarity than the background genes). For imprinted and pluripotency gene markers the scores were between 0.38 and 0.64 (P-values 0.006 to 4.5E-24) (Supplementary Figure B-2b). This strengthens our hypothesis that the identified imprinted lineage-specific genes play a role in the development of lineage cell states.

5.3.3 Large co-expressed module of imprinted, pluripotency, and hematopoietic genes

To characterize the relationship between imprinted, pluripotency and hematopoietic genes in diverging hematopoietic lineages and to gain insight into the structure of the underlying gene interaction network, we performed a combined (clustering) co-expression and functional analysis of the three gene lists. Interestingly, hierarchical clustering (HCL) analysis of the expression patterns of the 868 genes yielded one large core cluster (turquois) composed of 635 genes as well as four small clusters that occur along the diagonal of the heatmap (Figure 5-4a). We found that the core cluster contains 79% of all pluripotency genes (215 genes), 84% of the imprinted genes (73 genes), and 69% of the hematopoietic genes (319). This again supports an important role of imprinted genes in hematopoietic development and differentiation.

Next, we related the grouped genes of this core cluster to functional GO terms. The color-coded scatter plot in Figure 5-4b shows the ten most significant GO terms (after removing the child terms) that are enriched in this list of clustered genes. Some of the most significant terms are *cell differentiation, cellular developmental process, immune system development* and *hematopoiesis*. All biological processes listed in Figure 5-4b involve considerable numbers of imprinted, pluripotency, and hematopoietic genes. For instance, around 20% and 24% of the imprinted genes are involved in *cell differentiation* and *organ development,* respectively.

5.3.4 Putative transcriptional network involving imprinted genes

After finding such similarities in gene expression and functional association between major parts of the imprinted genes and pluripotency and hematopoiesis genes we asked how this interconnectivity may be established in the cell. In the simplest scenario, many imprinted genes would actually be part of the PluriNetWork or would belong to the hematopoietic genes, what is not the case. In our analysis, only five imprinted genes (*Gab1, Ins1, Phf17, Tsix,* and *Xist*) are present in the pluripotency list and three imprinted genes (*Axl, Calcr,* and *Gnas*) belong to the hematopoietic list. However the observed co-expression profiles might also be due to shared regulatory genes that control imprinted genes as well as pluripotency and hematopoiesis genes. Alternatively, imprinted genes might act as regulators of pluripotency and hematopoiesis genes. For these reasons, we generated an expanded imprinted gene network (IGN) that includes the annotated regulators of imprinted genes and genes that are regulated by imprinted genes (see methods). The IGN consists of 169 nodes and 1818 edges, each representing a direct interaction or regulation between two nodes. Hence, the IGN is highly interconnected, and this seems to be due to a specific functional module within the IGN (see next section) showing particularly high connectivity. Out of the 169 IGN genes, only 14 genes were not annotated on the microarray chip. Intriguingly, the IGN shares 32 genes (called IGN-shared genes) with either the pluripotency or hematopoietic genes; most of them are highly interconnected in the IGN network. 20 genes (*Ccnd1, Cdh1, Cdkn1a, Creb1, Gab1, Ins1, Myc, Mycn, Nfkb1, Phf17, Pou2f1, Oct4 (Pou5f1), Rela, Sp1, Sp3, Tert, Tgfb1, Tsix, Ube2i,* and *Xist)* are shared with the pluripotency genes and 17 genes (*Axl, Bcl2, Calcr, Cebpa, Egr1, Ets1, Gnas, Hoxa5, Jun, Junb, Myb, Myc, Nfkb1, Rara, Sp1, Sp3, Tgfb1*) are shared with the hematopoietic genes. Markedly, five genes (*Myc, Nfkb1, Sp1, Sp3,* and *Tgfb1*) appeared in both sets. Supplementary Table B-2 summarizes the complete gene sets considered in the analysis.

Fig. 4

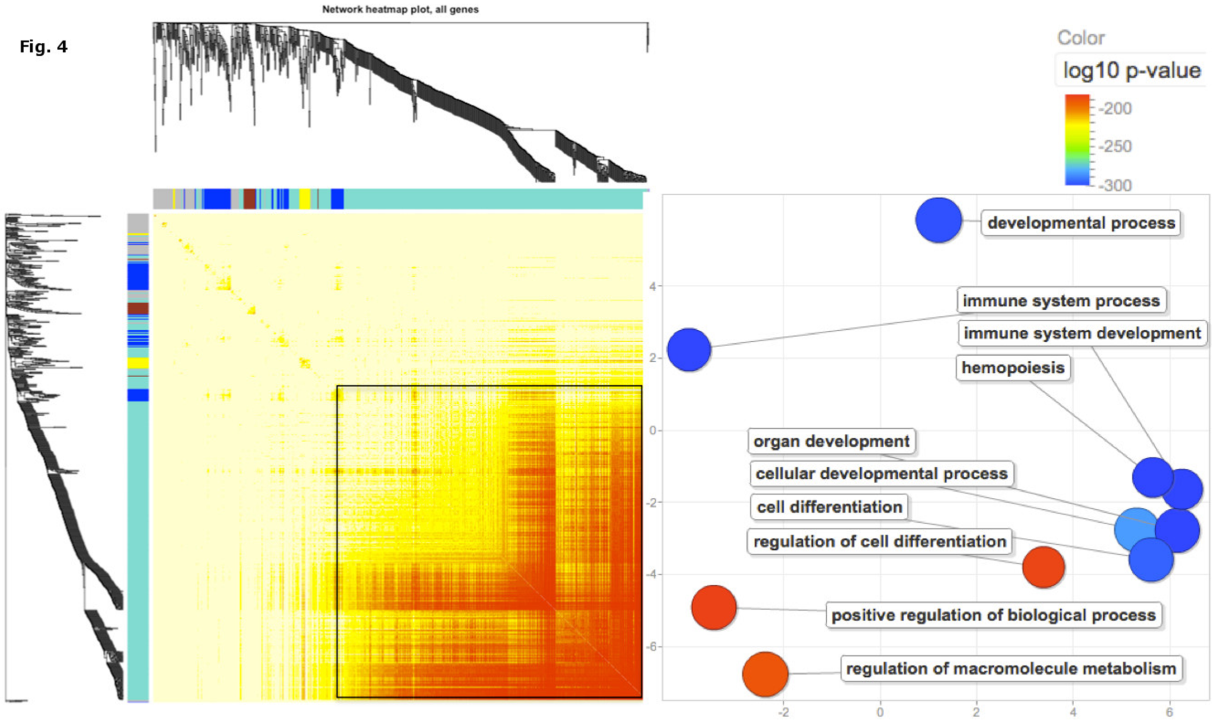


Figure 5-4 Co-expression and functional analysis of imprinted, pluripotency and hematopoietic genes.
A) On the left, heatmap depicting a gene interaction network based on the topological overlap matrix (TOM) [226] among the three gene lists. The TOM describes the distance between two genes in the co-expression network and reflects their similarity in terms of the commonality of the nodes they are connected to. A topological overlap of 1 between genes *i* and *j* implies that they are connected to the same genes, whereas a 0 value indicates that *i* and *j* do not share co-expression links to common genes. Each row and column of the heatmap corresponds to a single gene. Spots with light colors denote weak interaction and darker colors strong adjacency interaction. The dendrograms on the upper and left sides show the hierarchical clustering of genes. The turquoise, yellow, brown, blue, and grey colors represent the identified clusters and the black frame highlights the main gene cluster in turquois. B) Right, A scatterplot visualizing the top 10 enriched GO terms in the main (turquois) gene cluster in a two dimensional space of GO term semantic similarities. Node colors indicate the p-values for the enrichment of terms. The scatter plot was generated using the web tool REVIGO similarity [223]. This tool uses multi dimensional scaling to reduce the dimensionality of a matrix of the GO terms pairwise semantic similarity and projects the GO terms on the two axes. The axes thus visualize the semantic similarity of GO terms, but have no intrinsic meaning.

We then explored the network modularity according to the topological feature edge betweenness of the IGN network (Figure 5-5). Apparently, IGN-shared genes show a high modularity in IGN as they are grouped in only four modules. Importantly, 75% (24 genes) belong to a single module (Figure 5-5, green). By extracting the BP terms that are enriched in this gene module (excluding the 24 IGN-shared genes) we found many statistically significant functional terms related to differentiation and cell development such as GO:0008283 *“cell proliferation*”, GO:0009887 *“organ morphogenesis”,* GO:0030154 *“cell differentiation”*, and GO:0048863 *“stem cell differentiation”*.

5.4 Discussion

In this study we compared the expression patterns of 86 candidate imprinted genes and 272 pluripotency genes taken from the PluriNetWork [202] in ESC and during different stages of hematopoiesis to the global expression pattern of hematopoietic genes. We discovered that the three gene sets showed similar changes between pluripotent, intermediate, and differentiated stages suggesting that these gene sets have partially overlapping functions. Furthermore, we identified lineage markers from the three gene sets for three lymphoid and three myeloid branches that were found to exhibit significant functional similarities. In a collaborative fashion, they seem to participate in cellular differentiation and development in the corresponding lineages.

Interestingly, many imprinted genes shared very similar expression patterns with the pluripotency and hematopoiesis sets (Figure 5-2) similar to observations made for a smaller set of imprinted genes in murine HSCs [198]. This similarity was most pronounced for genes belonging to expression class 1 that are overexpressed in ESC. However, the functional similarity of imprinted genes and pluripotency and hematopoiesis genes, respectively, does not quite reach the level that is seen for the similarity between PluriNetWork and hematopoiesis genes. We attribute the observation that particular imprinted genes show a high variability of expression among the various stages of differentiation to the different roles played by these genes in the cell.

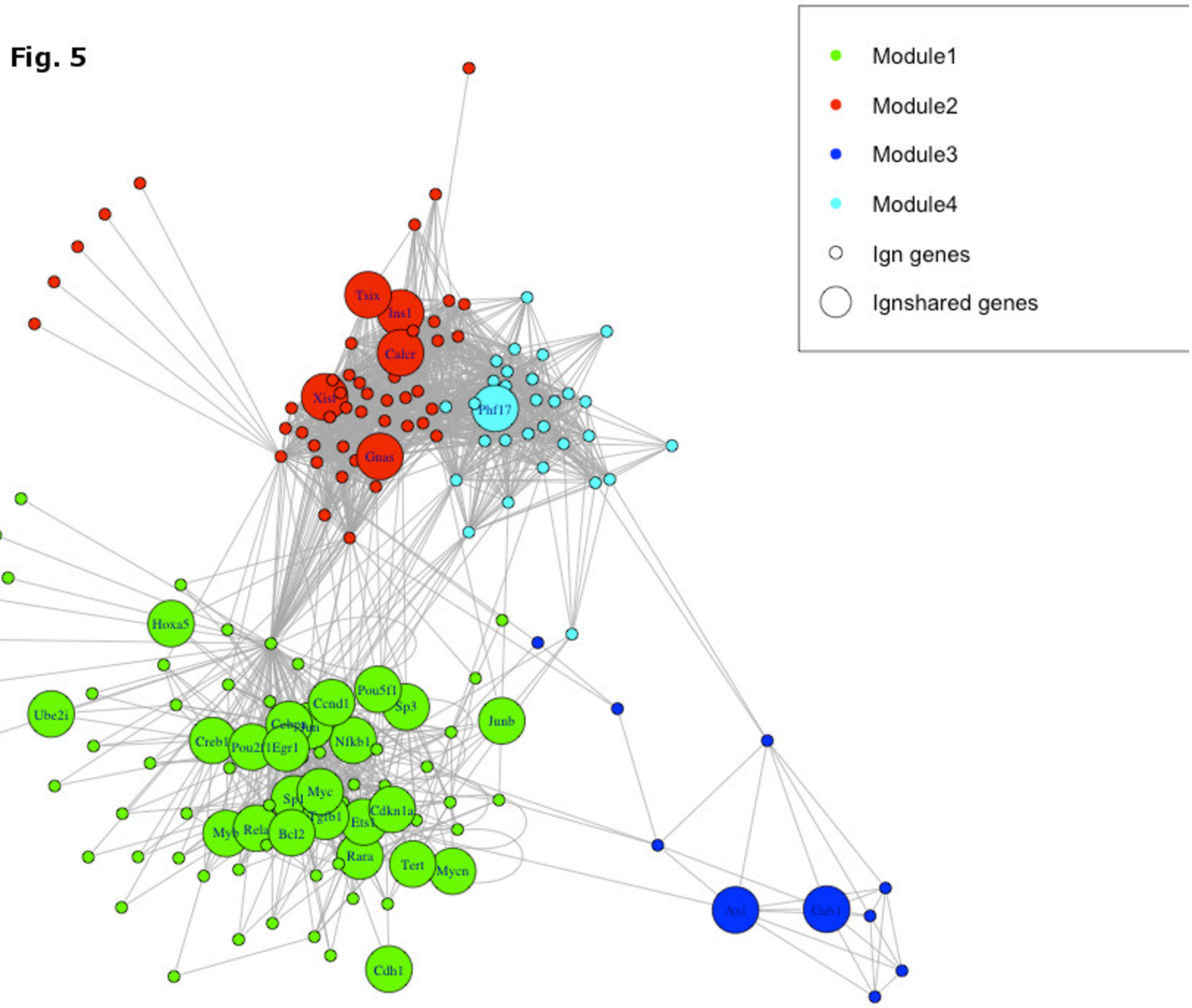


Figure 5-5 The expanded imprinted gene network (IGN) including all considered imprinted genes, transcriptions factors that regulate imprinted genes, as well as target genes regulated by imprinted genes. The edges of this graph may either indicate a significant degree of co-expression of two genes or a regulatory interaction. The IGN was clustered based on edge-betweenness. The full network is decomposed into only 4 topological modules. Large nodes represent the 32 genes that are shared with the PluriNetWork or with the set of hematopoietic genes. 24 of them belong to the main module (green).

Comparing the expression levels of known pluripotency with hematopoiesis genes showed that the compiled PluriNetWork [202] contains not only the GRN that keeps cells in the pluripotent state but appears also to be related to the regulation of the onset of cellular differentiation such as hematopoiesis. In fact, the GO terms *hematopoietic or lymphoid organ development, haemopoiesis, myeloid cell differentiation, leukocyte differentiation* are annotated to 44, 39, 27, and 23 pluripotency genes, respectively, suggesting that a significant portion of the pluripotency genes is indeed involved in hematopoiesis regulation. Moreover, Figures 1 and 2 demonstrate convincingly that the full set of pluripotency genes displayed pronounced variations during different stages of hematopoiesis and in individual cell lineages as well. These findings agree with previous studies that discussed the role of pluripotency genes in determining cell fate and controlling differentiation [227].

Different cell types showed pronounced differences in their gene expression profiles: most prominent was the high number of differentially expressed genes in B-cells and T-cells with a major contribution of pluripotency genes and to some extent also imprinted genes. This expression profile is interesting as a substantial portion of B-cells and T-cells serve as memory cells that can be induced by secondary infections to undergo further cell divisions. NK-cells that have recently been shown to have some potential for further cell divisions [228] tend also to have higher numbers of differentially expressed genes compared to the differentiated myeloids. Most of the lineage markers identified in this work were concordant with the findings of recently published studies. Generally, the ten lineage markers (*Cdkn1c, Ndn, Gatm, Phlda2, Air, Igf2r, Slc22a3, H13, Sfmbt2,* and *Peg12*) that participate in most lineages were demonstrated to be differentially expressed in the early onset of the hematopoietic process.[207, 229] More specifically, the identified erythrocytes lineage markers *Fli1, Mpl,* and *Gata2* were previously found to determine the erythrocytes signature [204, 230-232].

In order to validate that the identified lineage markers indeed have functional roles in the respective lineages, we have referred to their phenotypic gene knock-out characteristics documented in the MGI repository [225] (Supplementary Table B-1). Interestingly, knock-out of B-cell markers lead to abnormal B-cell differentiation and abnormal B-cell morphology etc.

In order to get more insight into the putative association between imprinted genes and their regulatory partners, we constructed an expanded IGN that is associated with genomic imprinting effects although only half of its genes are actually imprinted. Particularly, 32 IGN genes appeared in the pluripotency or hematopoietic sets. 24 of them (75%) belong to one topological module (Figure 5-5) suggesting that this module of 77 genes that are related to genomic imprinting due to their membership in IGN affects maintenance of pluripotency and hematopoietic differentiation in a cooperative manner. This module is also enriched in GO functional terms related to differentiation, development and hematopoiesis. Of the 32 IGN-shared genes *Oct4* and *Myc* are considered strategic players in maintaining the induced pluripotent state. Five IGN-shared genes (*Myc, Nfkb1, Sp1, Sp3,* and *Tgfb1*) were shared among the three sets indicating a regulatory role in cell differentiation and hematopoiesis. The TCF *Myc* belongs to the four known Yamanaka factors that play a significant role in cell reprogramming [233] and were shown to be sufficient for reprogramming differentiated cells into induced pluripotent stem cells [234]. *Myc* is believed to regulate expression of 15% of all human genes [235] and plays an important role in B-cell proliferation [236]. Recently, *Myc* and the changes in its expression level have been reported as a key player in embryonic stem cell development into megakarcocytes

[237], and in erythropoiesis [238]. *Tgfb1* is known to be involved in differentiation processes and was identified as a key regulator for HSCs homeostasis [239]. *Nfkb1*, when knocked out in mice, caused significant reduction in granulocytic progenitors and CFU-granulocytes [240] and it modulates proliferation and survival of erythroid progenitors derived from $CD34^+$ HSCs [241]. *Sp1* and *Sp3* control gene expression in myeloid cells [242] and during erythrocyte maturation [243]. Therefore, these 5 genes might be the major connectors between the IGN, pluripotency and hematopoiesis networks.

In summary, the present analysis suggested new aspects of the cellular regulation of the onset of cellular differentiation and during hematopoiesis. These involve, on the one hand, genes that were previously not discussed in the context of hematopoiesis and, on the other hand, involve genes that are related to genomic imprinting.

6. Application to breast invasive carcinoma

This chapter is a shortened version of the following publication:

- Mohamed Hamed, Christian Spaniol, Alexander Zapp, and Volkhard Helms, Integrative network based approach identifies key genetic elements in breast invasive carcinoma. BMC Genomics, 2015. 16 (Suppl 5): p. S2.

Synopsis

In this chapter, we demonstrate the usefulness of the integrative network-based approach to identify genetic key elements that could possibly drive the tumorogenesis in human breast cancer. The introduced approach was able to reveal strong associations between regulatory elements from four consistent genomic data sources: gene expression, DNA methylation, miRNA expression, and somatic mutations. Integrative screening of miRNAs, mRNAs, and genetic variations can contribute to an improved understanding of human diseases and hence to a better prognosis and treatment. Taken together, these findings endorse the reliability of the proposed approach so that it can be applied in a similar fashion to other cancer types, complex diseases, or for studying cellular functions where such multi-dimensional genomic datasets are available.

Abstract

Breast cancer is a genetically heterogeneous type of cancer that belongs to the most prevalent types with a high mortality rate. Treatment and prognosis of breast cancer would profit largely from a correct classification and identification of genetic key drivers and major determinants driving the tumorigenesis process. In the light of the availability of tumor genomic and epigenomic data from different sources and experiments, new integrative approaches are needed to boost the probability of identifying such genetic key drivers. We present here an integrative network-based approach that is able to associate regulatory network interactions with the development of breast carcinoma by integrating information from gene expression, DNA methylation, miRNA expression, and somatic mutation datasets.

Our results showed strong association between regulatory elements from different data sources in terms of the mutual regulatory influence and genomic proximity. By analyzing different types of regulatory interactions, TF-gene, miRNA-mRNA, and proximity analysis of somatic variants, we identified 106 genes, 68 miRNAs, and 9 mutations that are candidate drivers of oncogenic processes in breast cancer. Moreover, we unraveled regulatory interactions among these key drivers and the other elements in the breast cancer network. Intriguingly, about one third of the identified driver genes are targeted by known anti-cancer drugs and the majority of the identified key miRNAs are implicated in cancerogenesis of multiple organs. Also, the identified driver mutations likely cause damaging effects on protein functions. The constructed gene network and the identified key drivers were compared to well-established network-based methods.

The integrated molecular analysis enabled by the presented network-based approach substantially expands our knowledge base of prospective genomic drivers of genes, miRNAs, and mutations. For a good part of the identified key drivers there exists solid evidence for involvement in the development of breast carcinomas. Our approach also unraveled the complex regulatory interactions comprising the identified key drivers. These genomic drivers could be further investigated in the wet lab as potential candidates for new drug targets. This integrative approach can be applied in a similar fashion to other cancer types, complex diseases, or for studying cellular differentiation processes.

6.1 Background

Breast cancer is one of the most common and predominant cancer types that affects millions of cases and causes thousands of deaths every year [148, 244]. With an individual probability of 12% to develop breast cancer, it is the most frequently diagnosed cancer type among women and accounts for the second-highest number of fatalities (15%) of female cancer patients besides lung cancer [245]. Due to its complexity and heterogeneity [246], the molecular mechanism and regulatory patterns underlying breast carcinoma have not been completely unraveled so far.

Treatment and prognosis of cancer development relies largely on a correct classification of the histological grade and identification of the major determinants driving the tumorigenesis process. To better address this, many studies have attempted to build predictive models by analyzing and integrating heterogeneous data sources. For example, Cava et al. presented an effective discrimination of breast cancer types based on a support vector machine classifier combining copy number variations, SNP data, and the expression values of miRNAs, and mRNAs [247]. Also, miRNA-mRNA interactions were combined with transcription factor (TF)-gene interactions to unravel the combinatorial molecular regulations that facilitate progression of colorectal and breast cancer [118, 172]. Along the same lines, the integration of gene expression data with protein interaction networks into integrated weighted networks has already proven fruitful in a variety of applications within cancer genomics [248-263]. In general, the combination of microarray studies with mathematical models such as network theory allows to define relationships between genes and to discover interacting networks and pathways, improving the understanding of complex diseases [264].

In recent years, novel network-based approaches have been introduced to improve the understanding of complex human diseases from multiple perspectives. For instance, differential network analysis attempts to better characterize disease phenotypes under two different conditions by studying the changes in the related network interaction patterns [248, 249, 257, 258, 265-269]. In cancer genomics, the differential network approach was able to identify essential gene modules that lead to crucial novel biological insights and significant implications for understanding tumorigenesis [249, 257, 258].

In the light of the recent availability of tumor genomic data and the complexity of the related high throughput analysis, new integrative approaches are needed to boost the probability of successfully identifying the associated genetic key drivers, the causal regulators, the related mutations, biomarkers, and their molecular interactions that potentially drive tumorigenesis. To this end, this study presents an integrative network-based approach based on whole-genome gene expression profiling, DNA methylome, miRNA expression, and genomic mutations of breast cancer samples from the TCGA portal [66]. Based on this, we constructed a gene regulatory network that conforms to the conditions of such biological data and we identified network modules of dysregulated genes. Each module turned out to have distinct functional categories, cellular pathways, as well as oncogene and tumor suppressor specificity. We also extracted breast cancer specific subnetworks from the human genome regulatory interactome induced by the dysregulated miRNAs and the dysregulated mRNAs. Furthermore, we demonstrated a strong association between the different genetic molecules in terms of the interchangeable regulatory effect and genomic proximity. Then, we identified putative genetic key drivers/determinants from genes, miRNAs, and somatic mutations that could possibly drive the oncogenic processes in breast cancer.

Our findings are strongly supported by independent evidences. For instance, the protein products of about one third of the identified driver genes are known binding targets of anti-breast cancer drugs, and most of the identified key miRNAs are implicated in cancerogenesis of multiple organs. Moreover, all the identified driver mutations are predicted to cause damaging effects on structures and functions of the related proteins. The rest of the identified driver molecules represent novel potential candidates for new drug targets and further experimental research is warranted to confirm these findings.

6.2 Methods

See the integrative network-based approach presented in section 3.4.

6.3 Results and discussion

6.3.1 Differential analysis

We developed and applied an integrative network-based approach to conduct combinatorial regulatory network analysis in the context of breast invasive carcinoma with the aim of identifying the major genetic drivers that lead to tumorigenesis (Figure 3-9). We processed mRNA expression, DNA methylation, miRNA expression, and somatic mutation datasets for 131 tumor samples and 20 control samples of healthy tissues. The differential analysis of the mRNA expression, DNA promoter methylation, and miRNA expression data gave 1317 differentially expressed genes, 2623 differentially methylated genes, and 121 differentially expressed miRNAs, respectively.

6.3.2 TF-gene interactions

The expression profiles of the 1317 identified differentially expressed genes were used to compute the co-regulation strength between genes using the topological overlap (TOM) measure. Then, we performed hierarchical clustering (HCL) to construct the undirected co-expression network. HCL yielded 10 segregated network modules that contain between 26 and 295 gene members (Table 6-1). For the seven smallest modules, we collected the related directed regulatory interactions available in three online regulatory databases (JASPAR [83], TRED [82], and MSigDB [70]) and used them as a prior for a Bayesian learner to learn the causal probabilistic regulatory interactions and to generate a directed network topology, (see methods for details). The three largest modules (blue, brown, and turquoise) comprised too many nodes that exceeded the complexity that can be handled by the Bayesian learning approach. Hence, we deliberated the co-expression networks for these three modules by requiring a tighter co-expression threshold and used the obtained network modules for further analysis. It should be mentioned that the Bayesian approach prevents cyclic topology such as self-regulation, which is the case for many genes. Therefore, we note that self-regulatory interactions are not considered in this study. Next, the GRN network modules were pruned in order to maximize consistency between gene expression profiles, methylation fingerprints of gene promoters, and the inferred regulatory interactions. This helps to contextualize the network to the biological experiments from which it was reverse engineered. We removed 89 inferred interactions whose target genes are downregulated and their expression profiles showed absolute anti-correlation measure > 0.65 with their methylation profiles. In those cases we reasoned that downregulation of these target genes was most likely due to their promoter methylation and not due to TF binding [79].

By linking the network modules genes to GO and KEGG annotations via over representation analysis (ORA), we identified the most significant metabolic processes and functional categories that were enriched in each network module and showed relevance to breast cancer, see Table 6-1.

Table 6-1 The key driver elements identified from TF-gene interactions and miRNA-mRNA interactions. For the 10 gene modules identified in TF-mRNA interactions, we list counts of the involved genes, the most significant GO and KEGG terms, and the identified key driver genes from each module. Similarly for the miRNA-mRNA interactions, we list the key driver molecules of both genes and miRNAs. The driver genes, whose protein products are known to be targeted by drugs, are underlined and marked in red.

	Module	**Gene count**	**Top GO category**	**Top KEGG categories**	**Key driver count**	**Key drivers**
TF- mRNA interactions	black	41	Regulation of transcription	Pathways in cancer, Renal cell carcinoma	5	SORBS3, ZNF43, ZNF681, RBMX, POU2F1
	blue	247	Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	Cell cycle, Prostate cancer, Melanoma	9	AR, BRCA1, ESR1, JUN, MYB, RPN1, E2F1, E2F2, PPARD
	brown	195	Anatomical structure morphogenesis	Leukocyte transendothelial migration	5	TMOD3, CREB1, POU5F1, SP3, TERT
	green	110	Cellular macromolecule metabolic process	Endometrial cancer, Insulin signaling pathway	15	B4GALT7, OS9, CDC34, MAN2C1, MYO1C, SH3GLB2, INPP5E, PLXNB1, USF2, PPP1R12C, CDK9, DAP, E4F1, E2F4, USF1
	grey	148	Anatomical structure development	Sulfur metabolism	18	AHCTF1, NQO2, FGFR2, CCDC130, ABCG4, BIRC6, CA6, SP4, RNF2, SPRR1B, C16orf65, DNAJC5G, SNCAIP, GRIK5, SLC6A4, SMAD1, DAD1, POU4F2
	magenta	26	Regulation of metabolic process	p53 signaling pathway, Alzheimer's disease	3	ATF6, NGEF, POGK
	pink	30	Transcription initiation from RNA polymerase II promoter	Basal transcription factors	4	CCDC92, TMEM70, RNF139, E2F5
	red	93	Regulation of cellular process	Endometrial cancer, Neurotrophin signaling pathway	14	ATP1B1, STAT3, ABCB8, MYC, TGFB1, SP1, TP53, PCGF1, SUMF2, GTF3A, IPO13, GMPPA, HTR6, TGIF1
	turquoise	295	Regulation of cellular metabolic process	p53 signaling pathway, Pancreatic cancer, Apoptosis	2	UBL5, RNF111
	yellow	132	Immune system process	Chemokine signaling pathway, Natural killer cell mediated cytotoxicity	19	APOC1, CD2, CD79B, LRRC28, DAPK1, FAM124B, EML2, LAP3, TSPAN2, FCRL3, ELMO1, SLC7A7, RASSF5, SLC31A2, TRAF3IP3, GALNT12, ITGA4, SPI1, TFAP2A
	Total	1317				
miRNA- mRNA interactions	Genes	**Gene count**	**Top GO category**	**Top KEGG categories**	**Key driver count**	**Key drivers**
		869	Regulation of macromolecule metabolic process	Pathways in cancer, Pancreatic cancer, Prostate cancer	17	MYC, ATG4C, TGFB1, NFKB1, AKT1, EGR1, TP53, SOX10, SPI1, MECP2, E2F3, CREB1, TCF3, TPP1, FLICE, LPS, PACS1
	miRNAs	**miRNA count**	**Top functional categories**	**Top HMDD categories**	**Key driver count**	**Key drivers**
		120	miRNA tumor suppressors, immune response, Onco-miRNA , cell death, human embryonic stem cells regulation	Breast cancer (65), Neoplasms (58), Melanoma (56), Ovarian Neoplasms (51), Pancreatic Neoplasms (38), Prostatic Neoplasms (38)	68	mir-126, mir-609, mir-488, mir-191, mir-200c, mir-200a, mir-30a, mir-30d, mir-335, mir-190b, mir-223, mir-106b, mir-519e, mir-210, mir-379, mir-203, mir-205, mir-708, mir-29c, mir-29a, mir-182, mir-183, mir-127, mir-187, mir-425, let-7g, let-7d, mir-152, mir-155, mir-21, mir-22, mir-758, mir-921, mir-922, mir-375, mir-377, mir-181a-2, mir-657, mir-302d, mir-100, mir-10b, mir-10a, mir-625, mir-629, mir-92a-2, mir-26b, mir-25, mir-145, mir-143, mir-141, mir-221, mir-193b, mir-193a, mir-374a, mir-134, mir-146a, mir-31, let-7a-2, mir-27a, mir-27b, mir-133a-1, let-7i, mir-93, mir-23a, mir-148a, mir-196a-2, mir-487b, mir-149

For instance, the red and green modules are enriched with the endometrial cancer pathway, which is tightly associated with breast cancer and subsequent treatment [270]. Also, the magenta and turquoise modules were significantly involved in the p53 signaling pathway, a tumor suppressor gene showing one of the largest frequencies of SNPs among all human genes that have been related to cancer [148]. It has also important roles in diagnosis, in prognostic assessment and, ultimately, in treatment of breast cancer [271-275]. The inferred network topologies for the first three modules (red, green, and magenta) highlighting their identified driver genes are presented in Figure 6-1. Other network modules are shown in Figure C-1. Then we utilized the gplk solver [94] via OpenOpt [95] on the 10 inferred network modules to find the minimal set of nodes that dominate and regulate all nodes in each network. In total, we identified 94 key dominating/driver genes in all network modules (Table 6-1). The follow-up analysis of these driver genes is discussed below.

6.3.3 miRNA-mRNA interactions

To extract the breast cancer specific subnetworks from the human genome wide regulatory interactome induced by miRNAs and mRNAs, we examined two possible regulation types between the differentially expressed miRNAs and mRNAs: miRNAs regulating target mRNAs and mRNA products (TFs) regulating expression of the miRNAs. We relied on the experimentally validated interactions of both types in building the two networks, (see methods for details). The identified miRNA→mRNA interactions consist of 65 unique miRNAs and 770 unique genes involved in 1949 links. The TF→miRNA interactions include 112 unique TFs and 100 unique miRNAs composing 336 links. A total of 869 genes (including TFs) and 120 miRNAs were present in the combined miRNA→mRNA and TF→miRNA interaction network. 13 mRNAs and 45 miRNAs were common in both interaction types. The 869 genes were mostly involved in regulation of macromolecular metabolic processes and cancer pathways of multiple organs (Table 6-1). Moreover, the HMDD (Human miRNA Diseases Database) [133] analysis of the 120 miRNAs revealed their implication in cancerogenesis of various organs (Table 6-1). Next, the two networks comprising the dysregulated miRNAs and mRNAs as well as the interactions among them were combined and further analyzed using OpenOpt [95] and gplk solver[94] to identify genetic drivers and major regulators. This yielded in total 85 key dominating molecules (68 miRNAs and 17 genes) that regulate the entire network nodes (Table 6-1). The network topologies highlighting the dominating genes are shown in Figure 6-2.

Interestingly, some of the identified key driver genes such as *MYC, AKT1, and TP53* were previously implicated and significantly mutated in breast cancer samples [148]. Also the *TCF3* gene, a well-known TF controlling stem cell identity and self-renewal, is highly expressed in tumor samples and has a central regulatory role in the onset of breast cancer cell differentiation and tumor growth [276]. Additionally, many studies have reported the aberrant expression patterns of the *CREB1* gene and its role in breast tumor cell growth [277-280] suggesting its protein product as a worthwhile target for anti-cancer drugs [281, 282].

It has been demonstrated that the *E2F3* gene plays a critical role in the transcriptional activation of genes that control the rate of proliferation of tumor cells [149-151]. Furthermore, Vimala et al. [152] recently showed that the *E2F3* gene is overexpressed in 11 breast cancer cell lines and siRNA-*E2F3* based gene silencing facilitates the silencing of *E2F3* overexpression and limits the progression of breast tumors. This strongly conforms to our findings and implies that *E2F3* may be a potential therapeutic target for human breast cancer. HMDD analysis of the 68 driver miRNAs revealed that 36 miRNAs are involved in breast neoplasms, and the rest are associated with various cancer types such as hepatocellular carcinoma, adenocarcinoma, and prostate cancer. Also the identified key miRNA mir-29c as well as the key gene POU2F1 have recently been characterized as common hub nodes for three types of breast cancer [118]. Thus, unlike the traditional separate analysis of gene expression profiles [163-167] or the aberration of miRNA expression in cancer tissues [168-170], this integrated molecular analysis of the dysregulated miRNAs and mRNAs was able to uncover important aspects of the miRNA-mRNA interactome, the co-regulation mechanisms, and the underlying pathogenesis of human cancer.

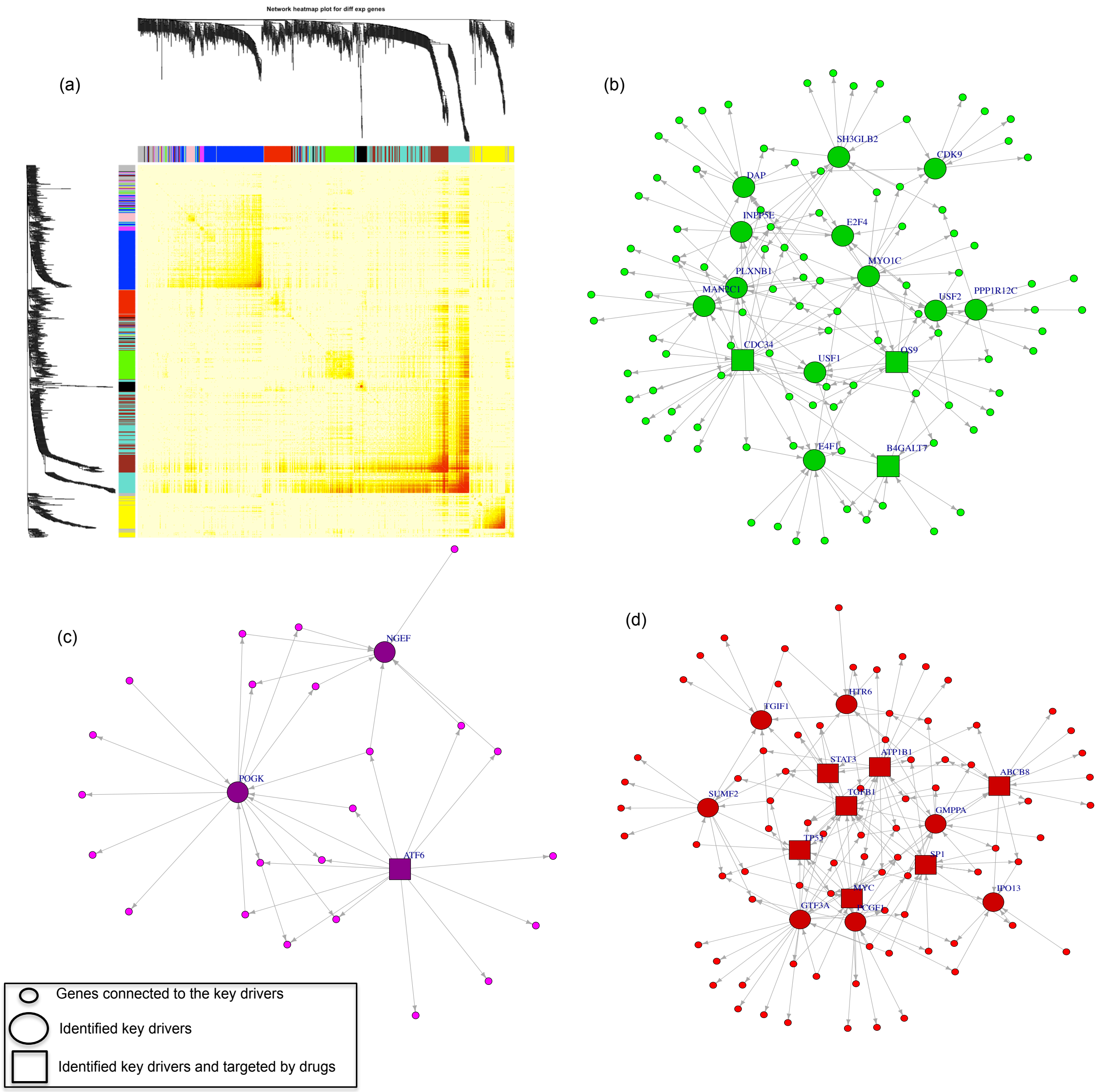


Figure 6-1 Gene network modules of TF-gene interactions.
(a) Topological overlap matrix (TOM) heatmap corresponding to the ten co-expression modules. Each row and column of the heatmap represent a single gene. Spots with bright colors denote weak interaction whereas darker colors denote strong interaction. The dendrograms on the upper and left sides show the hierarchical clustering tree of genes. (b), (c), and (d) are the final GRN networks highlighting the identified key drivers genes for the green, magenta, and red modules, respectively. Square nodes denote the identified driver genes that are targeted by drugs. Networks were visualized using the Igraph package in R.

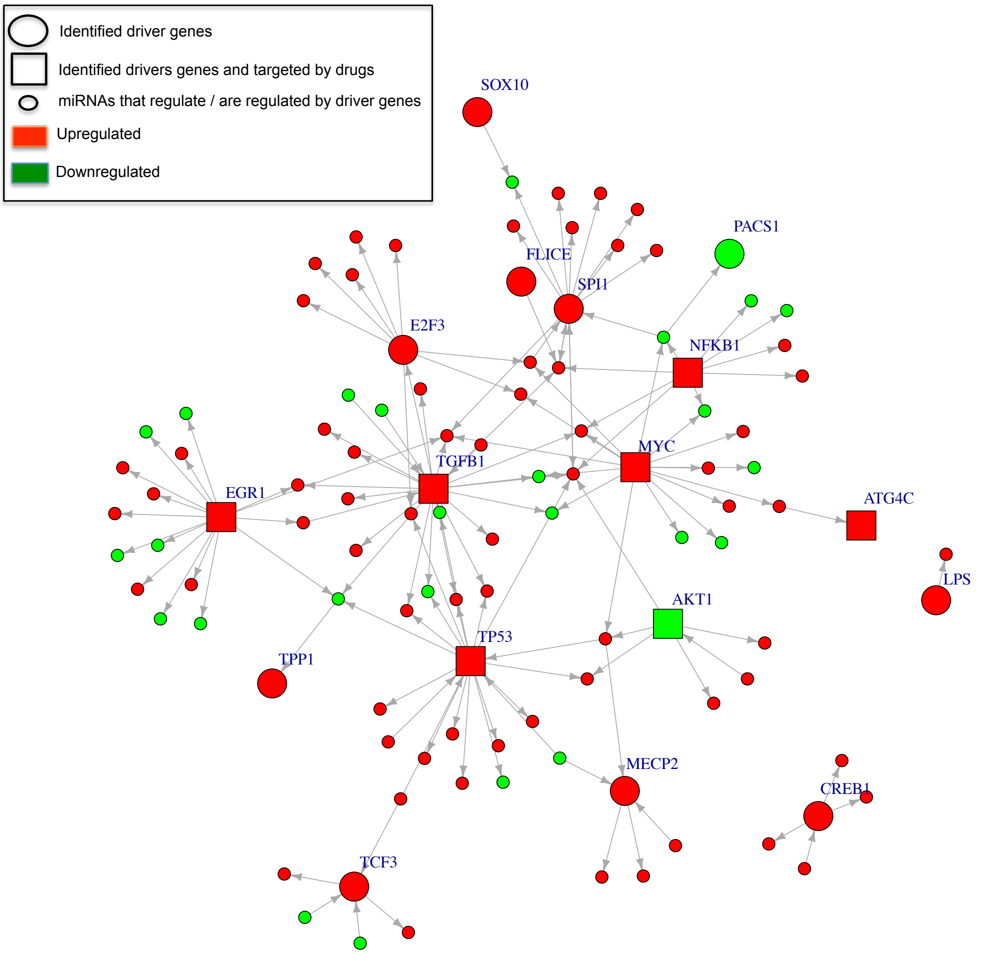


Figure 6-2 Regulatory interactions of the 17 key driver genes identified from miRNA-mRNA interactions. Large nodes represent key driver genes and small nodes represent miRNAs, which regulate or are regulated by these driver genes. Square nodes are the identified driver genes that are targeted by drugs. The network was visualized using the Igraph package in R.

6.3.4 Proximity analysis of somatic mutations

Although next generation sequencing of cancer genomes has unraveled thousands of DNA alterations, the functional relevance of most of these mutations and how they relate to other epigenetic mechanisms (such as DNA methylation and deregulation of miRNAs) are still poorly understood [100]. To this end, we scrutinized whether the significantly differentially expressed miRNAs are in genomic vicinity to the respective somatic variants so that dys-regulation of miRNA expression due to carcinogenesis may depend on the associated nearby somatic variants. We searched for the coding sequences of the dysregulated miRNAs in a genomic window of 250 kb around the somatic variants as previously described in [111]. We detected 21 cases of physical proximity between somatic variants and the deregulated miRNAs (Table 6-2), which are mostly located in chromosomes 1, 7, and 19 (Figure 6-3-a). These 21 cases encompass 15 distinct mutations and 20 distinct dysregulated miRNAs. To test the significance of these cases, we performed 1000 Wilcoxon tests against random SNV positions considering the same mutation frequency for each chromosome.

Table 6-2 The deregulated miRNAs in proximity to somatic mutations.
21 cases of miRNA-SNV pairs were identified. The genomic distance between miRNAs and SNVs is reported in base pairs. SNVs marked with (*) are the exclusive ones associated only with the dysregulated miRNAs and not with any of the non-dysregulated miRNAs.

miRNA	Chrom	SNP Position	SNP occurring gene	Genomic distance (in bp)
hsa-mir-181b-1	1	198711494 *	PTPRC	116508
hsa-mir-181a-1	1	198711494 *	PTPRC	116679
hsa-mir-1290	1	19186120 *	TAS1R2	37445
hsa-mir-9-1	1	156498803 *	IQGAP3	-108670
hsa-mir-205	1	209605636 *	MIR205HG	-158
hsa-mir-3129	2	189928732	COL5A2	69030
hsa-mir-145	5	148730786 *	GRPEL2	79423
hsa-mir-143	5	148730786 *	GRPEL2	77695
hsa-mir-106b	7	99662436 *	ZNF3	29180
hsa-mir-93	7	99662436 *	ZNF3	28955
hsa-mir-25	7	99662436 *	ZNF3	28747
hsa-mir-320a	8	22136963 *	PIWIL2	-34488
hsa-mir-199b	9	131048299	SWI5	-41299
hsa-mir-199b	9	131023779	GOLGA2	-16779
hsa-mir-152	17	46136186	NFE2L1	-21659
hsa-mir-520d	19	54254529	MIR522	-31179
hsa-mir-519e	19	54254529	MIR522	-71335
hsa-mir-1323	19	54254529	MIR522	-79307
hsa-mir-199a-1	19	10870471	DNM2	57631
hsa-let-7f-2	X	53644041	HUWE1	-59888
hsa-mir-718	X	153278098	IRAK1	7273

The deregulated miRNAs identified in the 21 cases were significantly closer to their somatic SNVs pairs in comparison to random SNV positions (p-value equals to 0.001). We also checked whether the non-dysregulated miRNAs (925 miRNAs) are in genomic proximity to the 15 somatic mutations involved in the 21 cases as well. We found that 52 non-dysregulated miRNAs (5.6%) were in vicinity to only 8 mutations so that the other 7 mutations are exclusively associated with the dysregulated miRNAs (Table 6-2).

Similarly, we analyzed the somatic mutations that mainly occur at differentially methylated CpG sites in promoter regions. Overall we identified 347 cases of SNV-differentially methylated gene pairs. These are mostly located on chromosomes 1, 5, and X (Figure 6-3-b). To address how changes in methylation levels caused by tumorigenesis correlate with mutation rates of different mutation genotypes, we separately analyzed the cases of up- and down-methylated genes. 234 cases involved up-methylated genes, whereas only 113 were associated with down-methylated genes. Generally, mutations in the promoter areas of up-methylated genes occur at a remarkably higher rate than its peers in down-methylated genes especially the C->T genotypes (Figure C-2) since methylated cytosines are prone to thymine transitions by via deamination. This result is in line with the findings of Xia et al. [21] who examined the relationship between DNA methylation and mutation rate. Further, we examined which of the above somatic mutations, which were identified on the basis of their vicinity to either dysregulated miRNAs or differentially methylated genes, could potentially drive tumor cell proliferation in breast cancer. For this, we applied the random forest as a machine learning method implemented in the CHASM tool [100] to distinguish between driver and passenger somatic mutations. As training set, we used the breast cancer labeled data (BRCA) curated from the COSMIC database [283] and provided by CHASM. We identified nine driver mutations (three from miRNA cases and six from differentially methylated gene cases) suggesting their causative role in breast tumorigenesis (Table 6-3). All these nine mutations are missense and lead to an amino acid substitution. Next, we analyzed the possible impact of the resulting amino acid substitution on structure and function of the respective protein using the PolyPhen [284] and SIFT [285] prediction tools. Interestingly, both methods predict damaging effects of these mutations on protein function conforming their role in driving cancer (Table 6-3).

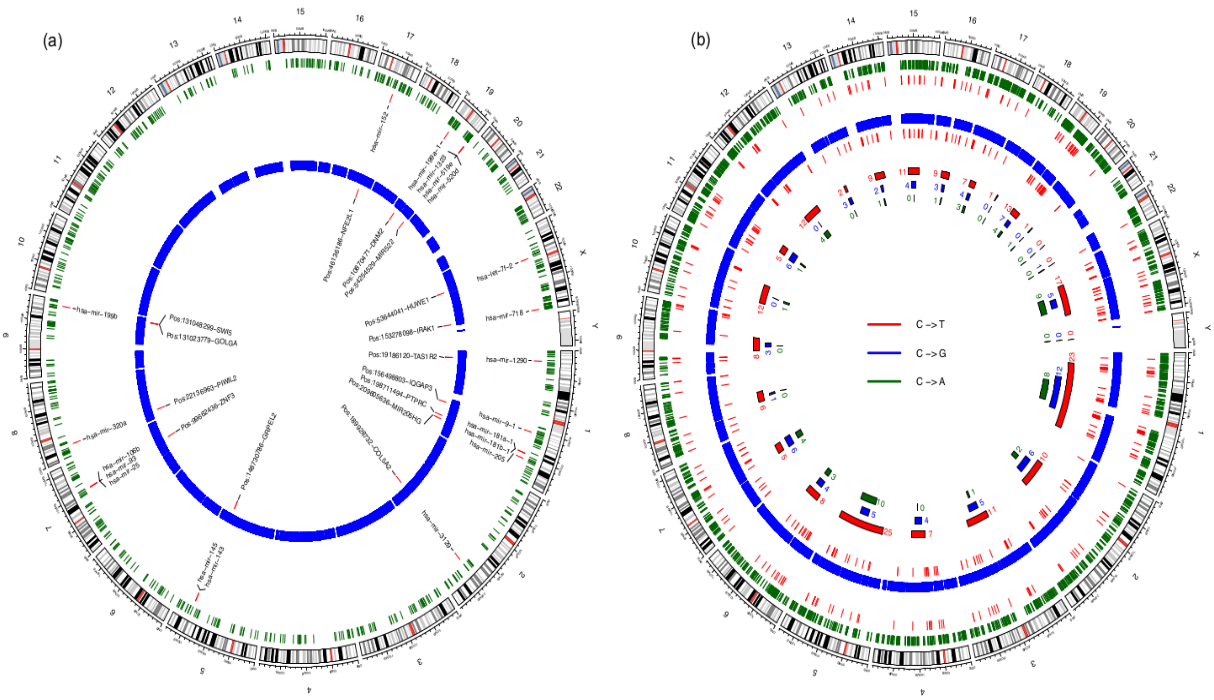


Figure 6-3 Proximity analysis of the somatic mutations with the dysregulated miRNAs and differentially methylated genes. Ideogram plots showing the genomic distribution for (a) the 21 cases of deregulated miRNAs adjacent to somatic mutations. The outer green circle shows the entire dataset of miRNAs , whereas the next highlighted red lines refer to the adjacent deregulated miRNAs (20 miRNAs where one miRNA is matched to 2 SNVs). The inner blue circle represent the entire set of somatic SNVs and the next highlighted red lines depict the SNVs matched to the 21 cases. (b) The 347 cases of somatic mutations occurring in the promoter regions of differentially methylated genes. The outer green circle shows the entire set of differentially methylated genes, whereas the next highlighted red lines refer to the identified cases adjacent to the somatic mutations. The inner blue circle represents the entire set of somatic SNVs and the next highlighted red lines depict the SNVs matched to the identified cases. The plot illustrates also the fractions of the three considered types of mutations (C->T, C->G and C->A) showing the occurrence frequency for each one.

6.3.5 Druggability analysis of protein products of the identified driver genes

As mentioned above, we identified 94 driver genes from the TF-mRNA interactions and 17 driver genes from the miRNA–mRNA interactions. The five well-known breast cancer associated genes *CREB1, MYC, TGFB1, TP53,* and *SPI1* were common in both sets. Hence, in total 106 driver genes were identified. Also, we characterized 68 dominating miRNAs from the miRNA-mRNA interactions, and nine driver mutations from the proximity analysis.

To identify driver genes marked as anti-breast cancer drug-targets, we looked up the drugs and the anti-neoplastic agents that target the proteins corresponding to the 106 driver genes based on the experimentally validated drug-targets reports (see methods). We found that 31% (33 proteins) of the proteins belonging to the identified driver genes are binding targets of at least one anti-breast cancer drug (Table C-2). These 33 genes are highlighted as square nodes in the network visualizations of TF-mRNA interactions (Figure 6-1, and Figure C-1) and miRNA-mRNA interactions (**Figure 6-2**). The remaining 73 driver genes were involved in the regulation of biological processes as well as metabolic processes of cancerogenesis in multiple organs such as lung, prostate, and bladder (Table C-1). This supports the hypothesis that products of the remaining 73 identified driver genes as well as the identified 68 driver miRNAs and the 9 driver mutations may open up new avenues for novel therapeutic drugs.

Table 6-3 List of the identified driver mutations ordered by CHASM score.
The CHASM score is defined as the fraction of trees in the Random Forest that voted for the mutation being classified as a passenger. Lower scores increase the confidence of driver mutations. P-values are calculated based on the null score distribution. The table reports also the changes in the related codons and amino acids. The SIFT and PolyPhen scores refer to the prediction of whether an amino acid substitution affects the function and structure of the human proteins. The SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences (lower scores represent high impacts), whereas the PolyPhen prediction uses physical and evolutionary comparative considerations (higher scores represent high impact and severe influence on the protein function and structure).

Chrom	Occurring gene	SNV position	CHASM score	P-value	Ref	Alt	Amino acids	Codons	SIFT score	PolyPhen score
1	PTPRC	198711494	0.158	6.00E-04	G	A	E/K	Gag/Aag	Deleterious (0)	probably_damaging (0.999)
8	TNKS	9413850	0.162	6.00E-04	C	T	S/F	tCc/tTc	Deleterious (0.01)	Unknown (0)
X	GRIA3	122319694	0.298	0.0119	C	A	F/L	ttC/ttA	Deleterious (0)	probably_damaging (0.996)
5	PCDHB14	140604126	0.308	0.0134	C	T	S/L	tCg/tTg	Deleterious (0.02)	Benign (0.368)
X	HUWE1	53644041	0.31	0.0136	C	A	R/L	cGa/cTa	Deleterious (0)	probably_damaging (1)
17	NFE2L1	46136186	0.326	0.0175	C	T	S/F	tCc/tTc	Deleterious (0.01)	probably_damaging (0.994)
9	NAIF1	130829249	0.336	0.0204	C	G	K/N	aaG/aaC	Deleterious (0)	probably_damaging (0.995)
2	KLHL23	170592167	0.354	0.0251	C	G	R/G	Cga/Gga	Deleterious (0)	probably_damaging (0.999)
12	KCNA1	5021107	0.384	0.0406	C	T	T/M	aCg/aTg	Deleterious (0)	probably_damaging (0.997)

6.3.6 Network validation and performance assessment

In order to validate the proposed approach and the constructed network topology [TF-gene interactions only], we applied a peer knowledge-based differential network method, KDDN (Knowledge-Guided Differential Dependency Network) [286] on the same dataset. The same prior was used for KDDN. The networks predicted by our approach showed 61% edges overlap with the inferred differential KDDN interactions due to tumorigenesis.

To assess the reliability of our predictions of key drivers, we further included another differential network method, DiffCoEx (Differential Co-expression Modules) [268] for identifying differential co-expression modules between two biological cohorts. As mentioned above, 33 genes (31%) out of the total 106 driver genes suggested here are known key drivers and are targeted by currently known drugs. In contrast, only 114 KDDN genes (~20%) out of 584 hot spot genes involved in the KDDN network, are binding targets for anti-cancer drugs (Figure 6-4).

We detected an overlap of 44%, and 16% of the key genes identified by our approach and those obtained by KDDN and DiffCoEx, respectively. DiffCoEx yielded five different modules of genes in which the correlation of gene pairs within the module was significantly different between normal and tumor samples (Figure 6-5). Only 151 genes (17%) out of total 886 genes involved in these modules were marked as anti-cancer drug targets. These percentages strongly support the reliability and robustness of our strategy in identifying genomic drivers that could be further experimentally examined as drug targets.

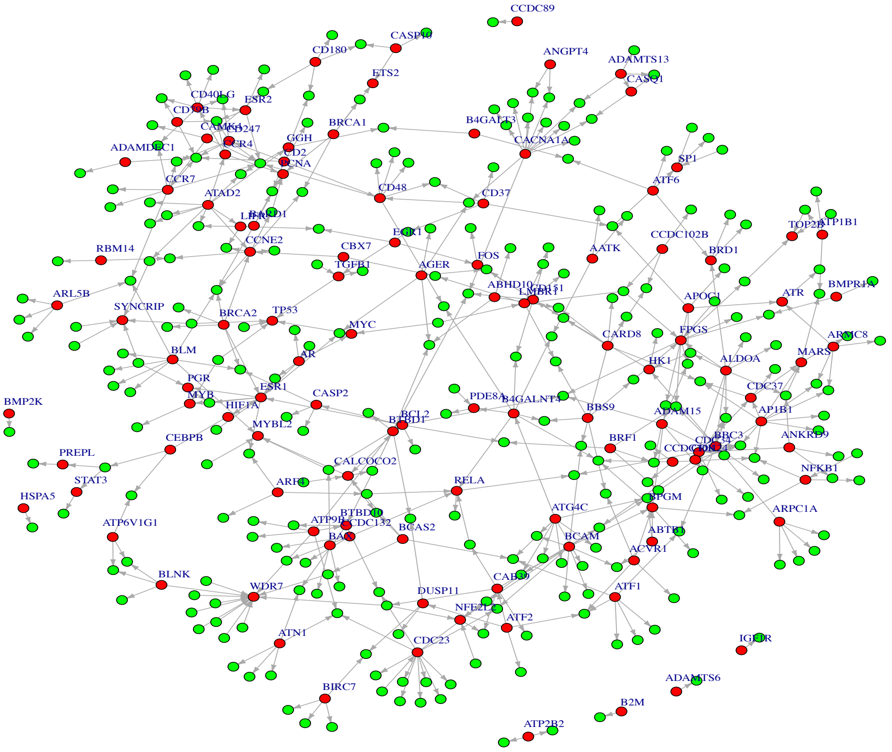


Figure 6-4 The network inferred using the KDDN method.
For clarity, we visualized only the known drug target genes (red and labeled) and the genes connected to them (green).

6.4 Conclusions

The enormously increasing availability of transcriptomic and epigenomic data from different biological experiments allow for deep and comprehensive integrative analysis. To this end, this study provides new insights into the complex regulatory mechanisms between gene expression, miRNA biomarkers, epigenetic modifications (represented at the level of DNA methylation) and genetic variants that are associated with the human breast cancer network.

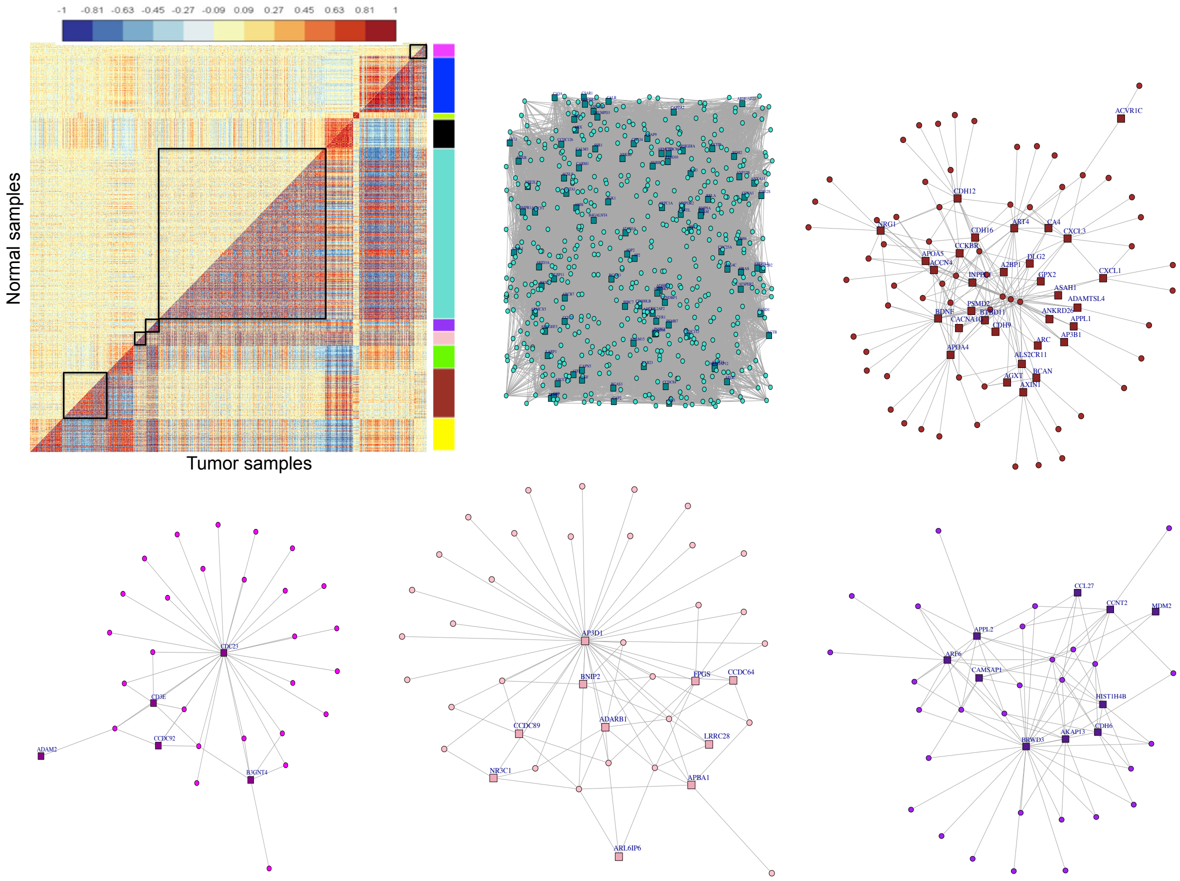


Figure 6-5 The network modules inferred using the DiffCoEx method.
Each network corresponds to the highlighted module color in the heatmap. For clarity, we visualized only the known drug target genes (labeled and square nodes) and the genes connected to them

In this work, we demonstrated an integrative network-based approach to conduct combinatorial regulatory network analysis and to identify genomic driver elements that control breast carcinomas. Our results showed a strong association between the regulatory elements of the heterogeneous data sources in terms of the mutual regulatory influence and genomic proximity. By analyzing three different types of interactions, TF-mRNA, miRNA-mRNA, and proximity analysis of somatic variants, we were able to identify various key driver elements (106 genes, 68 miRNAs, and 9 mutations) that could possibly drive breast invasive carcinomas. We also unraveled underlying regulatory interactions among these key drivers and other genetic elements in the breast cancer network. Interestingly, anti-breast cancer drugs target protein products of about one third of the key driver genes and most of the identified key miRNAs are involved in cancerogenesis of multiple organs. Also, the identified driver mutations are predicted to cause damaging effects on protein functions and structures.

These results expand our knowledge base of prospective genomic drivers and provide encouraging support that many of the novel identified genetic elements are potential targets for new drugs. We note that these key drivers were identified based on the presented computational framework and further wet lab work is warranted to confirm their efficacy as putative anti-cancer drug targets. Especially when combined with experimental validation, this network-based approach could promote novel insights on cancer genomic data to develop new therapeutic strategies and thus better treatment. Finally, this approach can be applied to other cancer types or complex diseases and could be extended for studying cellular development as well.

7. Application to diabetes in mouse

This chapter is a shortened version of the following publication:

- Irhimeh M.R, Barthelmes D, Mohamed Hamed, Zhu L, Helms V, Gillies M.C, Shen W, Novel Gene Regulatory Network in diabetic bone marrow-derived endothelial progenitor cells [In revision].

Synopsis

Differential network analysis concept has been recently introduced to improve understanding of cellular interactions of specific tissues and complex diseases. This chapter discusses the molecular mechanisms by which diabetes impairs Bone marrow-derived endothelia progenitor cells (EPC) in mouse using the differential network analysis approach that makes use of the implemented GRN pipeline. We note that all bioinformatics analysis in this study was performed by the author of this thesis.

Abstract

Endothelial progenitor cells (EPCs) are a group of rare cells that originate from bone marrow (BM) or the wall of blood vessels. They are believed to play an important role in the repair of injured vascular endothelial cells and assisting in reperfusion of ischemic tissue. Decreased production and/or loss of function of EPCs are associated with diabetic vascular complications such as diabetic retinopathy, nephropathy and cardiovascular disease. However, the molecular mechanisms by which diabetes impairs EPCs remain unclear. In this study we conducted microarray analysis of the differential gene expression between Akita diabetic mice and age-matched non-diabetic controls in BM-derived Lin$^+$ cells and Lin$^-$/VEGF-R2$^+$ EPCs isolated from animals 18 weeks after diabetes. EPCs were isolated using MACS technology based on hematopoietic lineage depletion followed by enrichment for VEGF-R2$^+$ cells to produce Lin$^-$/VEGF-R2$^+$ EPCs. Lin$^+$ fractions were kept and used as non-hematopoietic cells for analysis. RNA was extracted, processed and then hybridized to mouse WG-6 V2 beadchips, followed by data analysis. In total, 11 differentially expressed genes were identified as specific to BM EPCs including 3 genes (*CLCNKA, PIK3C2A, PTF1A*) with known association with diabetic complications and 8 genes classified as transcription factors (*PPARG, PPARA, VDR, FOXO1, AR, NFKB1, HNF4A, SREBF1*). Further analysis led to establishing a novel gene regulatory network specific to diabetic EPCs, which includes 11 main well documented diabetic genes and 47 genes and transcription factors regulating/regulated directly by those genes. Our results suggest that diabetes may influence specific signature genes in BM EPCs altering their capacity to proliferate and differentiate.

7.1 Introduction

Chronic diabetes is associated with endothelial cells (ECs) injury, cells forming the inner lining of blood vessels [287]. Such injury is believed to be repaired by resident endothelial cells, which have limited regenerative capacity [288, 289], resident endothelial progenitor cells (EPCs) [290, 291], and BM derived EPCs [292-294]. It has been reported that diabetes is associated with impairment of EPC function [295-297]. Diabetic patients were shown to have reduced EPC numbers in the peripheral blood (PB) [298, 299] and the ability of EPCs isolated from PB of people with diabetes to proliferate, form tubes and adhere *in vitro* is impaired [300, 301]. Most importantly, EPCs from diabetic individuals are less effective in repairing vascular injuries [300, 302, 303]. Several studies suggest that reduced number and/or dysfunction of EPCs in cell mobilization, proliferation, adhesion and incorporation into the vasculature may contribute to diabetic vascular complications [298, 300, 304].

Recently, our collaborators reported an impaired mobilization capacity of mouse BM Lin$^-$/VEGF-R2$^+$ EPCs in diabetic mice [305]. EPCs are usually defined based on their surface markers and proliferative and clonogenic potential and they are believed to be lineage and functionally heterogeneous [290]'[293]. It has been suggested that an insult to the stem cell niche might initiate or contribute to reduction in the numbers and impairment of EPC function [306]. These EPCs play an important role in regenerating the endothelium through migration, proliferation, differentiation and by secreting pro-angiogenic cytokines [307]. BM Lin$^-$/VEGF-R2$^+$ EPCs express VEGF-R2 and CD34 and they do not express CD31, CD45, CD14 and CD115. They have typical EPCs properties such as formation of cobblestone colonies, Dil-acLDL uptake, lectin binding and can incorporate into damaged blood vessels *in vivo* after intravitreal transplantation in eyes subjected to the laser-induced retinal vascular injury[293, 308] in association with differential expression of only two genes (*SDF-1 (CXCL12) and SELE*) in diabetic Lin$^-$/VEGF-R2$^+$ EPCs [305].

The majority of molecular studies on the impairment of diabetic EPCs function have been conducted on human EPCs isolated from the PB of people with a long history of diabetes. Thus, little is known about the changes occurring in EPCs located within the BM in the early stages of diabetes. In this study we isolated Lin$^+$ cells and Lin$^-$/VEGF-R2$^+$ EPCs from Akita diabetic mice and age-matched non-diabetic controls. In order to explore the molecular mechanisms by which early diabetes impairs the function of BM-EPCs, we conducted microarray analysis to profile differential gene expression and their regulatory interactions between diabetic and non-diabetic animals using well-established data analysis methods [208, 309].

7.2 Methods

7.2.1 Animals

The Akita mouse carries a dominant point mutation in the Insulin 2 gene on chromosome 7 resulting in the development of diabetes at approximately 4 weeks after birth with almost 100% penetrance. As female mice develop diabetes more slowly and less stably compared with males, only male mice heterozygous for the Ins2^{Akita} allele (diabetic group) as well as male mice homozygous for the wild type Ins2 allele (non-diabetic mice) were used in this study. Once diabetes was established (blood glucose level>13.3mmol/L), mice were monitored weekly for changes in bodyweight and blood glucose levels for 18 weeks. The blood glucose level was measured using Accu-Chek Performa (Roche, Germany). No supplemental insulin was given. Only mice with blood glucose levels consistently ≥ 13.3 mmol/L were used in this study. Nine diabetic mice and age-matched non-diabetic controls were used in this study.

7.2.2 Group design and comparisons

Four experimental groups were established: 1) Lin$^+$ cells from non-diabetic mice, 2) Lin$^+$ cells from diabetic mice, 3) Lin$^-$/VEGF-R2$^+$ EPCs from non-diabetic mice and 4) Lin$^-$/VEGF-R2$^+$ EPCs from diabetic mice. The Lin$^+$ cells were used as an internal reference to identify differential gene expression occurring not exclusively in Lin$^-$/VEGF-R2$^+$ cells. Six different comparisons were conducted between the four groups (Table 7-1). This setup allowed us to distinguish differential gene expression which specifically occurred in diabetic BM derived Lin$^-$/VEGF-R2$^+$ EPCs from that occurring in other phenotypes of hematopoietic lineage committed BM cells. Hence, only significant changes in gene expression observed in diabetic versus non-diabetic Lin$^-$/VEGF-R2$^+$ progenitor cells that did not occur in the Lin$^+$ population were considered in the final analysis.

7.2.3 Data processing

Raw expression values were background corrected, log$_2$ transformed and quantile normalized using the lumiR package [310] of the Bioconductor suite [220]. Expression profiles of redundant probe sets were merged by computing the mean of all probes related to single genes as reported before in [309]. Before the differential analysis, we removed the 25% of the genes that showed the least variability across the sample groups. Genes with higher variation were considered as potentially good candidates to be differentially expressed [221].

Table 7-1 The six possible comparisons (1-6) between the 4 groups of samples and the significance of each comparison to the study analysis.

Comparison	Compared groups	Significance/meaning
1	Non-diabetic Lin$^+$ vs diabetic Lin$^+$	Effect of diabetes on Lin$^+$
2	Non-diabetic Lin$^+$ vs Non-diabetic EPC	Difference between Lin$^+$ and EPC genes in healthy conditions
3	Diabetic Lin$^+$ vs diabetic EPC	Difference between Lin$^+$ and EPC genes in diabetic conditions
4	Non-diabetic EPC vs diabetic EPC	Effect of diabetes on EPC
5	Lin$^+$ vs EPC	Difference between Lin$^+$ and EPC combined
6	Non-diabetic vs diabetic	Difference between non-diabetic and diabetic cells

7.2.4 Differential expression analysis

The six comparisons of samples were compared by differential expression analysis using three methods: 1) Significance Analysis of Microarray (SAM) [90], 2) moderated t-test,[221] 3) the area under the curve of the receiver operator characteristics (AUC ROC) [221]. Genes that were classified as differentially expressed genes by at least two of those three methods were included in the list of differentially expressed genes. We focused on genes that are exclusively involved in the fourth comparison (non-diabetic Lin$^-$/VEGF-R2$^+$ EPCs vs diabetes Lin$^-$/VEGF-R2$^+$ EPCs) and not in any of the other comparisons.

7.2.5 Gene regulatory network (GRN)

We applied the GRN pipeline presented in section 3.2 on the expression data of the differentially expressed genes for the healthy and diseases samples separately. Then we used the concept of the differential network analysis to infer the statistically significant topological changes in transcriptional networks representing the two biological samples (non-diabetic Lin$^-$/VEGF-R2$^+$ EPCs vs diabetes Lin$^-$/VEGF-R2$^+$ EPCs). This gave a differential network of 109 genes (including 25 differentially expressed genes and 84 TCFs) and 347 edges. Figure 7-1 depicts the differential network analysis approach employing the GRN pipeline for data processing and constructing the diabetes-specific grn for BM derived EPC cells.

We downloaded a list of 266 diabetic-associated genes from Mouse Genome Informatics (MGI) database [225]. Out of these 109 genes, only 11 genes (3 differentially expressed genes and 8 TCFs) belong to the diabetic associated gene list. The full list of genes in the GRN is provided in supplementary Table D-1. To extract only the network module related to the onset of diabetes, we removed unconnected nodes and considered only regulation links where either the "FROM" or "TO" nodes belong to the 11 diabetic genes identified above. Expression heat maps and PCA plots were generated by R [188]. The GRN network was visualized using the igraph package in R.

7.2.6 Functional enrichment

The functional enrichment and annotation analysis was conducted as reported before in [208]. Briefly, enriched KEGG Pathways and GO functional categories were identified using the DAVID tool [135]. We determined which pathways / functional terms were annotated to at least 2 genes and were statistically overrepresented in the study gene set against the full mouse genome (control). Enrichment was evaluated through the hyper-geometric test using a p-value threshold of 0.05.

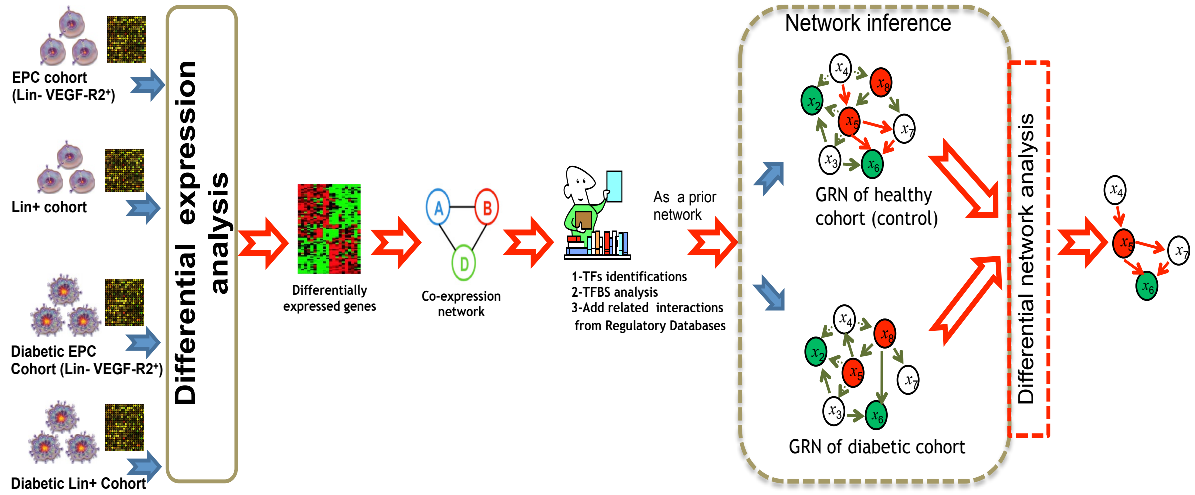

EPC cohort (Lin- VEGF-R2+)
Lin+ cohort
Diabetic EPC Cohort (Lin- VEGF-R2+)
Diabetic Lin+ Cohort
Differential expression analysis
Differentially expressed genes
A
B
D
Co-expression network
1-TFs identifications
2-TFBS analysis
3-Add related interactions from Regulatory Databases
As a prior network
Network inference
x4
x8
x2
x5
x7
x3
x6
GRN of healthy cohort (control)
GRN of diabetic cohort
Differential network analysis

Figure 7-1 The differential network approach utilizing the GRN pipeline.

7.3 Results

7.3.1 Probes summarization and filtration

One of the aims of this study was to identify genes that are differentially expressed between 2 sample groups (diabetic and non-diabetic Lin$^-$/VEGF-R2$^+$ EPCs). Microarray probes of mouse WG-6V2 beadchip (45,281 probes) were summarized by considering the mean of the expression values of all probes related to each gene in each sample. This yielded at the end 30,869 mouse genes instead of 45,281 probes. Then, non-specific pre-filtering was performed removing the 25% of all genes showing the least variability across the two sample groups before the differential analysis.

7.3.2 Differential expression analysis

To identify differentially expressed genes between non-diabetic and diabetic samples, only the summarized 30,869 genes were used for comparisons. After applying the three differential expression methods (SAM, moderated t-test, and AUC ROC) on each of the six comparisons, genes that were identified as differentially expressed by at least two out of the three methods were chosen for further analysis.

Then, using Venn diagrams we identified those genes which were exclusively included in the fourth comparison (between non-diabetic EPCs and diabetic EPCs) and not in any of the other comparisons) as shown in Figure 7-2. This process identified 80 genes that are specific to comparison 4 only.

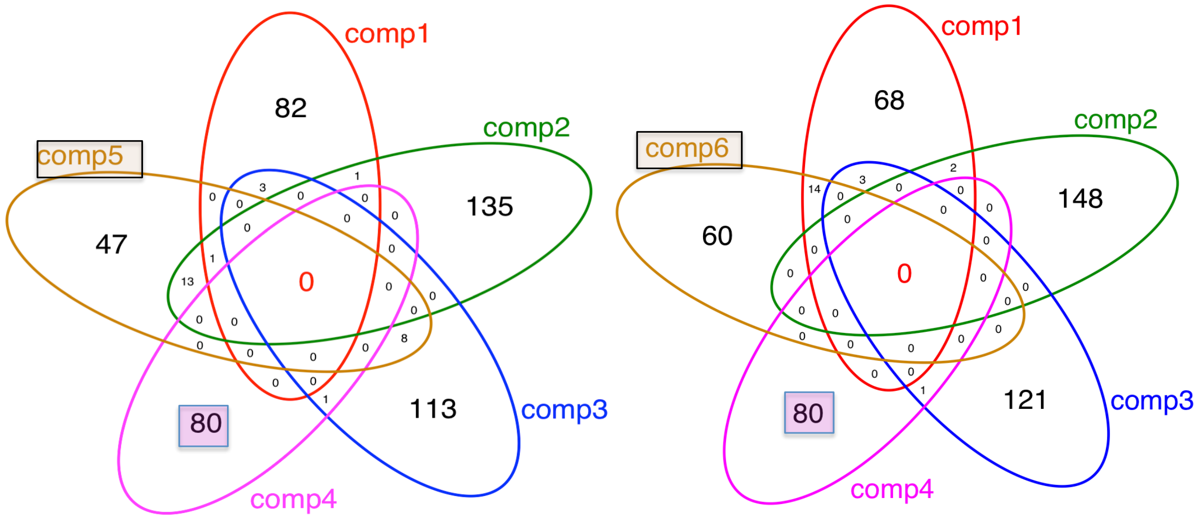


Figure 7-2 Venn diagrams showing overlapping differentially expressed genes among the 6 comparisons. See Table 1. (A) Comparisons 1-5, (B) comparisons 1-4 and 6. In both Venn diagrams the same 80 genes were found specific to comparison 4 (non-diabetic EPCs versus diabetic EPCs).

This approach allowed us to identify those genes which might influence BM EPCs but not hematopoietic lineage committed cells by targeting certain regulatory elements. To further investigate changes in diabetic associated genes in BM EPCs, 266 diabetic associated genes were downloaded from MGI[225] and then cross matched with the 80 identified genes. We found that the identified differentially expressed genes were significantly associated with the list of diabetic associated genes (p-value 0.0319 using hyper-geometric test) because the three genes *CLCNKA* (down-regulated) and *PTF1A* and *PIK3C2A* (both up-regulated) were common between the two lists. A heat map was generated to show the relative gene expression among the four sample groups (Figure 2A). Then we selected non-diabetic and diabetic EPCs groups to generate a heat map for the relative expression of the 80 identified genes (Figure 7-3). To show how the differentially expressed 80 genes are separated between non-diabetic EPCs and diabetic EPCs, principle component (PCA) analysis was conducted. The PCA clustered the differentially expressed genes into down-regulated and up-regulated genes based on their relative expression levels.

7.3.3 Gene Regulatory Network

Compiling the GRN for the identified differentially expressed genes revealed 84 TCFs that were regulating 25 out of the 80 differentially expressed genes. Figure 7-4 shows an expanded GRN with 25 + 84 = 109 genes including the identified TCFs. Only 8 out of the 84 TCFs (*PPARG, PPARA, VDR, FOXO1, AR, NFKB1, HNF4A, SREBF1*) were classified as diabetes related genes. Thus, a total of 11 (3 differentially expressed genes and 8 TCFs) diabetic related genes were present in the expanded GRN network of 109 genes. Interestingly, the hyper-geometric test conducted on the list of all the 109 GRN genes showed a highly significant association with the diabetes related genes (P-value 1.138

e-09). When considering only genes and TCFs that are connected to the 11 diabetic related genes, a final GRN module, which includes 58 nodes that could potentially drive and dissect the early diabetes and related dysfunctions in BM EPC cells, was compiled and visualized in Figure 7-5. The details of the final 58 GRN genes and TCFs are listed in supplementary Table D-1.

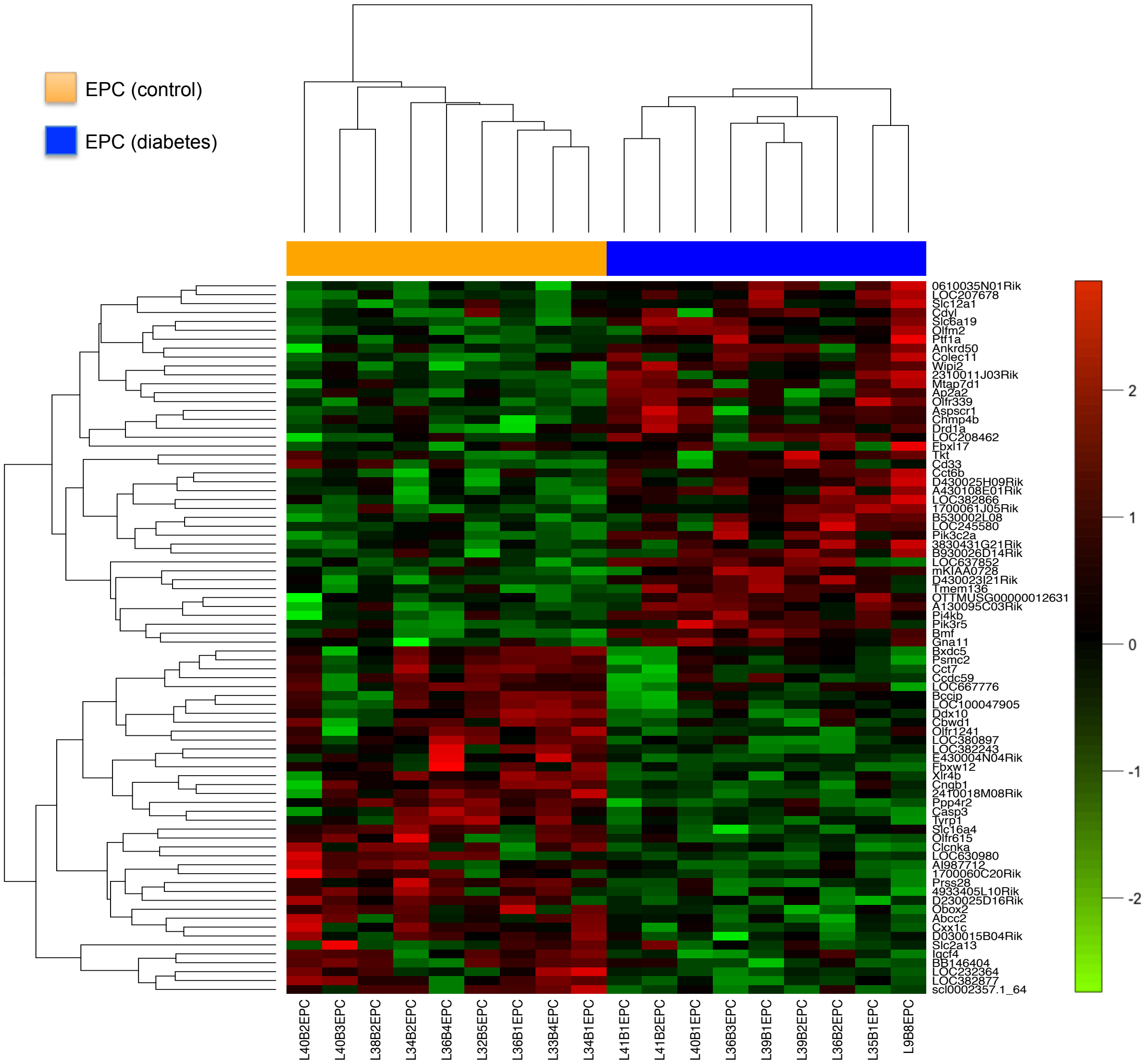


Figure 7-3 Heat maps of the microarray analysis results.
Differentially expressed 80 core enrichment genes in comparison 4 (non-diabetic EPCs versus diabetic EPCs). Green spots represent down-regulated genes, and red spots represent up-regulated genes. The order of genes is obtained by hierarchical clustering. The orange color represents the non-diabetic EPCs while the blue color represents the diabetic EPCs.

7.4 Discussion

Previous studies have investigated EPCs in various diabetic complications. Although the methods used have been quite different and subsets of the investigated EPCs were also disparate, they all found significant dysfunction of diabetic EPCs.[290, 293, 305] Numerous explanations for the dysfunction of diabetic EPCs have been proposed, including increased oxidative stress, NADPH oxidase activation, an altered nitric oxide pathway and increases in inflammatory cytokines [311]. However, it is unlikely that the dysfunction of diabetic EPCs could be explained by a single independent mechanism when diabetes is known to be a complex patho-physiological syndrome that leads to EPCs dysfunction and subsequently vascular damage at several levels. Thus we used a microarray analysis approach and complemented that with powerful and well-established data analysis methods [208, 309] to investigate genes and TCFs that are potentially affected in diabetic EPCs and could be responsible for their dysfunction. We were able to construct a novel gene regulatory network specific to BM EPCs that have been exposed to short period of diabetes.

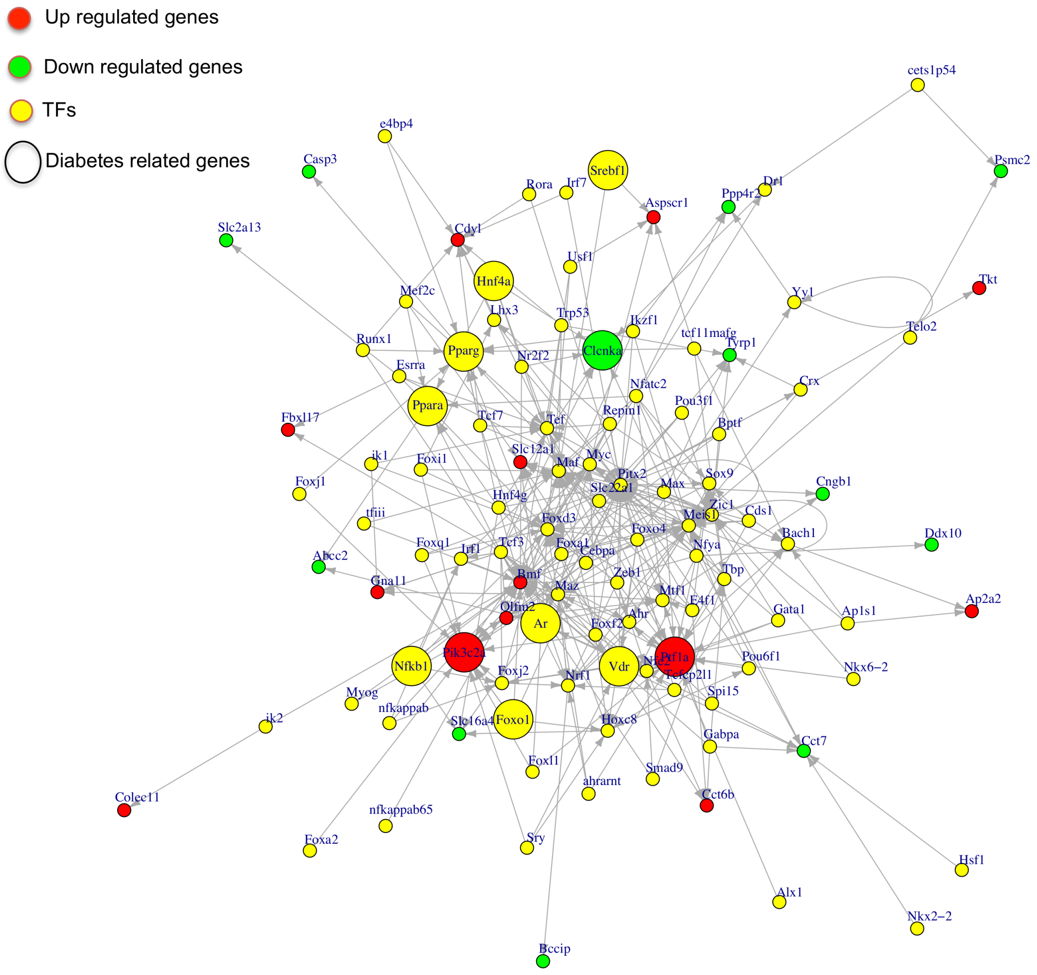


Figure 7-4 Gene regulatory network common to EPCs.
The expanded diabetes gene regulatory network in EPCs including 109 genes (25 differentially expressed genes and 84 transcription factors that regulate them).

We previously demonstrated that BM Lin$^-$/VEGF-R2$^+$ EPCs form cobblestone colonies in culture, express surface markers such as VEGF-R2 and CD34, and are more primitive than other described EPCs with a limited capacity to participate in vascular repair. It appears that EPC function in the early stages of diabetes (18 weeks) is impaired, in particular their ability to mobilize, rather than their ability to proliferate, leading to trapping of EPCs in BM [293, 305, 312]. Since the exact mechanism underlying this impaired mobilization is still unknown, identifying the responsible genes through the use of high throughput methods such as microarray may lead to valuable insights into the pathogenesis of diabetic vascular disease. Most microarrays contain probes for many more genes than are differentially expressed.

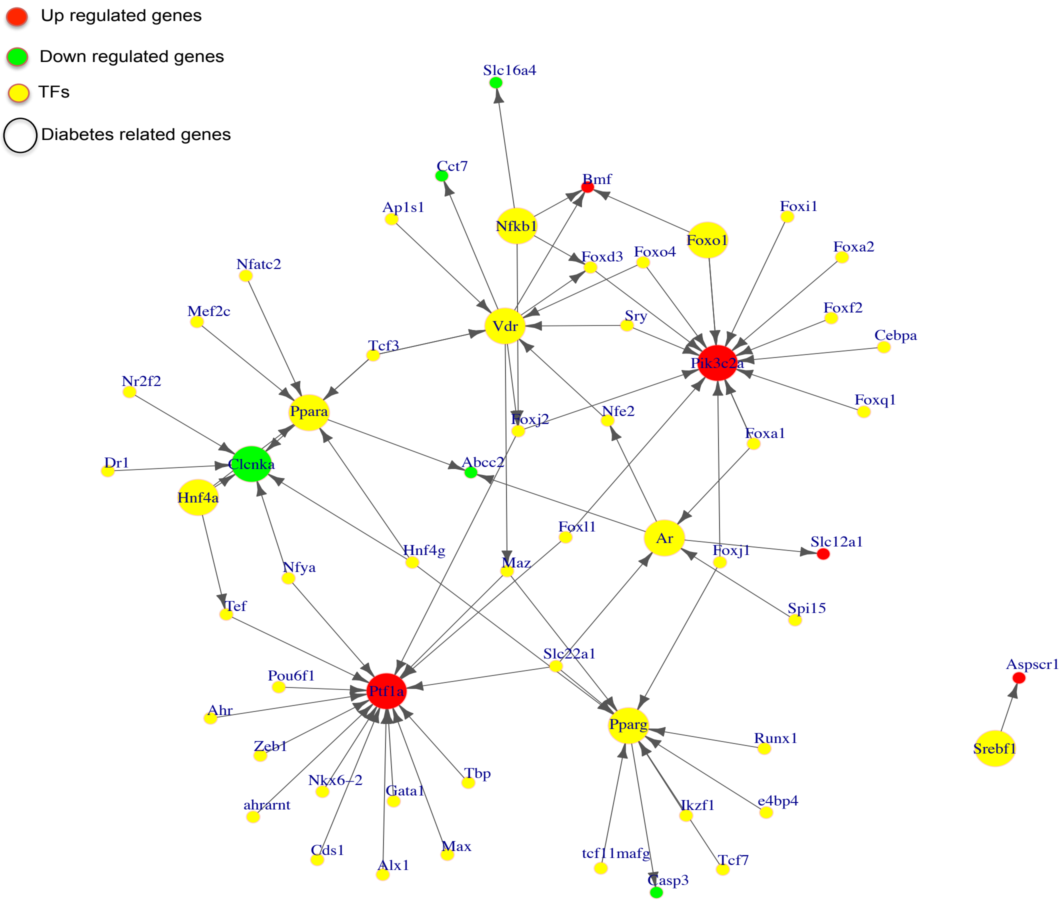


Figure 7-5 A final gene regulatory network (GRN) module of the identified 11 diabetes-related genes and the regulatory elements directly connected to them.

To alleviate the loss of power from the formidable multiplicity of gene-by-gene hypothesis testing, we carried out a non-specific (done without reference to the parameters or conditions of the tested RNA samples) pre-filtering step [309]. This helped us remove from consideration a set of probes/genes that are not differentially expressed under any comparison. We found it most useful to select genes on the basis of variability [313]. Only the genes that show a noticeable variation across samples can potentially be differentially expressed among our groups of interest. Usually, in microarray analysis the coefficient of variation is used to filter the probes, then a threshold is chosen (*i.e.*, 0.1) and all genes with coefficient of variance below the threshold are removed from the analysis. In this study we applied instead a non-specific filter based on the variance itself and removed the 25% genes showing the lowest variability across the samples, as previously described [208, 309].

Following the analysis of the differentially expressed genes, KEGG analysis of the 58 genes of the final GRN module revealed highly significant pathways (Table 7-2 A) such as 'maturity onset diabetes of the young' and 'peroxisome proliferator-activated receptor (*PPAR*)' signaling pathways. The *PPAR* pathway plays a critical role in the regulation of diverse biologic processes. There are 3 main isotypes of *PPAR* gene (*α, β* and *γ*). In our GRN two TCFs/genes (*PPARα* and *PPARγ*) are found as major players in the diabetic EPCs network. Previously, *PPARα* has been implicated in the hepatic metabolic response to diabetes mellitus. *PPARγ* is expressed in all major cells of the vasculature (*e.g*., endothelial and smooth muscles cells) and there mutations lead to severe insulin resistant and type-2 diabetes [314]. More recently, *PPARα* was found to play an important role in the regulation of EPC trafficking. Activation and over-expression of *PPARα* both suppressed EPCs mobilization and homing induced by hypoxia, which was shown to be through the inhibition of the *HIF-1α/SDF-1* pathway [315]. This supports our finding where *PPARα* was found to be up regulated in diabetic EPCs, consistent with an anti-angiogenic role.

Another TCF that is part of the diabetic EPC network is *FOXO1* which plays a crucial role in regulating gluconeogenesis and glycogenolysis by insulin signaling [316]. *FOXO1* regulates *PIK3C2A*, one of the major diabetic genes in the GRN, which encodes for the PIK3C2A enzyme (PI3K family) that is activated by insulin. Thus, in diabetic conditions the activity of PIK3C2A enzyme is expected to be suppressed, which may result in upregulation of *FOXO1*. Both *PIK3C2A* and *PTF1A*, which are the only two differentially upregulated diabetic genes in the GRN, are found to be regulated by other identified genes and TCF in the GRN. In other words, they do not regulate any of the identified GRN genes and TCF leading to the assumption of them being the main and the most important genes of the diabetic GRN and perhaps they interact directly with the system (Figure 7-5).

The other mechanism by which *FOXO* regulates diabetic EPCs is via the oxidative stress activated *P66SHC-AKT-FOXO* pathway [317]. *P66SHC* was reported to be involved in EPC dysfunction due to hyperglycemia. When *P66SHC* was deleted in mice, the BM-derived EPCs showed increased survival and more resistance to oxidative stress [318]. Based on Figure 5 *FOXO1* is linked directly with *PIK3C2A*. This may contribute to the dysfunction of EPCs observed in diabetes through a negative effect of hyperglycemia-induced oxidative stress on the *PIK3C2A/FOXO1* axis and the activation of *P66SHC-AKT-FOXO* pathway. Another list of 26 highly significant functional terms that are relevant to our cell type and injury were identified through GO analysis of the 58 genes of the final GRN module (Table 7-2 B).

There is strong evidence that supports the concept of diabetes altering the number of circulating EPCs [319, 320], which are likely trapped in the BM, and impairing their vasoreparative potential resulting in premature senescence [298, 321]. In this study there was an obvious dominance of pathways that involve insulin and glucose metabolism, secretion, response and regulation (13 pathways) and progenitor cells and epithelial cells differentiation, proliferation and development (13 pathways).

Table 7-2 Selected highly significant (A) KEGG and (B) GO terms and the GRN genes that are involved in each category.

A)

KEGG Subcategory name	p-value	Number of genes	Gene IDs of test set in subcategory
Pathways in cancer	8.82E-07	9	*AR FOXO1 RUNX1 PPARG TCF7 CEBPA NFKB1 MAX CASP3*
Acute myeloid leukemia	4.76E-05	4	*RUNX1 TCF7 CEBPA NFKB1*
Maturity onset diabetes of the young	8.65E-05	3	*HNF4A HNF4G FOXA2*
MAPK signaling pathway	2.18E-03	5	*NFATC2 MEF2C NFKB1 MAX CASP3*
Adipocytokine signaling pathway	2.35E-02	2	*PPARA NFKB1*
Phosphatidylinositol signaling system	3.09E-02	2	*CDS1 PIK3C2A*
PPAR signaling pathway	3.38E-02	2	*PPARA PPARG*
Apoptosis	4.14E-02	2	*NFKB1 CASP3*

B)

GO Subcategory name	p-value	Number of genes	GeneIDs of test set in subcategory
Cell differentiation	3.19E-12	21	*TCF3 HNF4A MEF2C IKZF1 AR VDR RUNX1 FOXJ1 PPARG FOXA1 FOXO4 CEBPA FOXD3 FOXA2 SRY ALX1 NKX6-2 ZEB1 NR2F2 CASP3 PTF1A*
Regulation of cell proliferation	3.52E-07	10	*HNF4A AR FOXO1 FOXJ1 PPARG TCF7 FOXO4 CEBPA ZEB1 CASP3*
Regulation of cell differentiation	5.93E-08	10	*IKZF1 AR VDR FOXJ1 PPARG FOXA1 CEBPA FOXA2 NKX6-2 ZEB1*
Cell fate commitment	1.61E-11	9	*TCF3 MEF2C AR PPARG FOXA1 FOXA2 NKX6-2 CASP3 PTF1A*
Negative regulation of cell Proliferation	2.46E-08	8	*HNF4A AR FOXJ1 PPARG FOXO4 CEBPA ZEB1 CASP3*
Epithelium development	4.48E-07	8	*AR VDR PPARG FOXA1 FOXA2 ALX1 ZEB1 CASP3*
Positive regulation of cell differentiation	8.51E-06	6	*IKZF1 PPARG FOXA1 CEBPA FOXA2 NKX6-2*
Hemopoiesis	3.23E-05	6	*TCF3 IKZF1 RUNX1 FOXJ1 CEBPA ZEB1*
Epithelial cell differentiation	4.39E-07	6	*AR PPARG FOXA1 FOXA2 ZEB1 CASP3*
Negative regulation of cell Differentiation	8.33E-04	4	*FOXJ1 FOXA2 NKX6-2 ZEB1*
Blood vessel development	2.61E-03	4	*MEF2C FOXO1 RUNX1 NR2F2*
Response to insulin stimulus	5.18E-04	3	*PPARA FOXO1 SREBF1*
Response to glucose stimulus	5.92E-05	3	*HNF4A SREBF1 CASP3*
Regulation of cell cycle	9.67E-03	3	*HNF4A FOXO4 CASP3*
Glucose metabolic process	3.31E-03	3	*PPARA HNF4A FOXO1*
Glucose homeostasis	9.52E-05	3	*HNF4A FOXA1 ASPSCR1*
Cellular carbohydrate metabolic process	2.67E-02	3	*PPARA HNF4A FOXO1*
Wnt receptor signaling pathway	4.15E-02	2	*TCF7 FOXL1*
Regulation of insulin secretion	6.19E-03	2	*HNF4A SREBF1*
Regulation of glucose metabolic process	1.58E-03	2	*HNF4A FOXO1*
Positive regulation of glucose metabolic process	2.15E-04	2	*HNF4A FOXO1*
Positive regulation of Gluconeogenesis	9.90E-06	2	*HNF4A FOXO1*
Monosaccharide biosynthetic process	2.84E-03	2	*HNF4A FOXO1*
Insulin secretion	1.19E-02	2	*HNF4A SREBF1*
Insulin receptor signaling pathway	2.97E-03	2	*FOXO1 SREBF1*
Cellular response to insulin stimulus	6.19E-03	2	*FOXO1 SREBF1*

Thus, it is likely that diabetes influences the expression of EPC genes that are specific to those pathways causing impairment in their vasoreparative potential. Based on the MGI database [225] *eNOS* (*NOS3*), *SDF-1*, *CXCR4*, and *SELE* are all specific EPC genes yet they were not differentially expressed in any of the six comparisons in this study. However, we found that all of them were regulated by two TCFs that are identified in diabetic EPCs GRN (*USF1* and *NFKB1*). *SELE* appeared to be directly regulated by *NFKB1* while *CXCR4* is directly regulated by *USF1*, whereas *SDF-1* and *eNOS* were indirectly regulated by *USF1*. Thus, *USF1*, and *NFKB1* might be driving the expression changes of those genes during diabetes (Figure 7-6).

Dysfunction of eNOS signaling has also been implicated in EPC dysfunction in diabetes. The dysfunction has been linked with decreased *eNOS* activity [322, 323] and the *eNOS* deficient (*NOS3*$^{-/-}$) mouse had impaired EPC mobilization and angiogenesis [322]. The expression and phosphorylation of *eNOS* are essential for the survival, migration and angiogenesis facilitated by EPCs and ECs [324, 325]. Human EPCs that overexpress *eNOS* have increased migratory potential, increased ability to incorporate into tube-like structures and to differentiate into endothelial spindle-like structures [326]. We did not observe a significant change in *eNOS* expression in diabetic EPCs. Nevertheless, two of the genes that were found to be differentially expressed in diabetic EPCs (*NFKB1* and *USF1*) regulate *eNOS* indirectly, which could explain previous reports. We have previously reported that *eNOS* expression in BM Lin^{-}/VEGF-R2^{+} progenitor cells was very low indicating that they are early progenitor cells [312] since late EPCs have higher expression levels of eNOS [327].

Although there are many reports that diabetes causes reduction in PB EPC number [300, 328], others have reported an increase in EPC number in the circulation in specific animal models [329] while we did not find any significant effect of diabetes in mice [308]. We previously reported down regulation of *SDF-1* and *SELE* genes in diabetic EPCs.[312] Since EPCs have the ability to produce SDF-1 [330] and SDF-1/CXCR4 is a known EPC mobilization and maturation axis [331], this down regulation of *SDF-1* may contribute to the impaired mobilization of diabetic EPCs. Thus the observed decrease of diabetic EPCs in PB could be attributed to the impaired mobilization ability from BM to PB leading to EPC BM-trapping and not to the impaired proliferation. In this study neither *SDF-1* nor *SELE* were found differentially expressed but were found to be closely regulated by two important diabetic genes *NFKB1* and *USF1*. Circulating EPCs have the ability to express *SELE* [332], a sign of EPC activation [333]. We observed previously a 2.5 fold increase in the expression of *SELE* in diabetic EPCs. The up-regulation of *SELE* in diabetic EPCs may be attributed to increased production of interleukin 1 and tumor necrosis factors caused by the diabetic condition [334].

EPCs have a direct role in angiogenesis [288, 335]. We found that the same gene (*USF1*) that regulates *eNOS* indirectly through *ETV4* and *ESR1*, regulates *SDF-1* via *ESR1* as well, which is implicated in diabetes [336]. Thus *ESR1* may have a direct role in impaired diabetic EPCs. We also found that *USF1* regulates *CXCR4* directly, which indicates that *eNOS*, *SDF-1* and *CXCR4* are all closely related and regulated by the same key gene(*USF1*). Since *ETV4* is a downstream target gene of *FGF* signalling pathway which promotes tumour growth and angiogenesis [337], then such involvement in angiogenesis could explain its role in EPC, which under diabetic conditions have reduced potential to migrate, proliferate and form tubes [338]. Leading to the conclusion that *USF1* is affected by diabetic environment and could be responsible for EPC angiogenic activity.

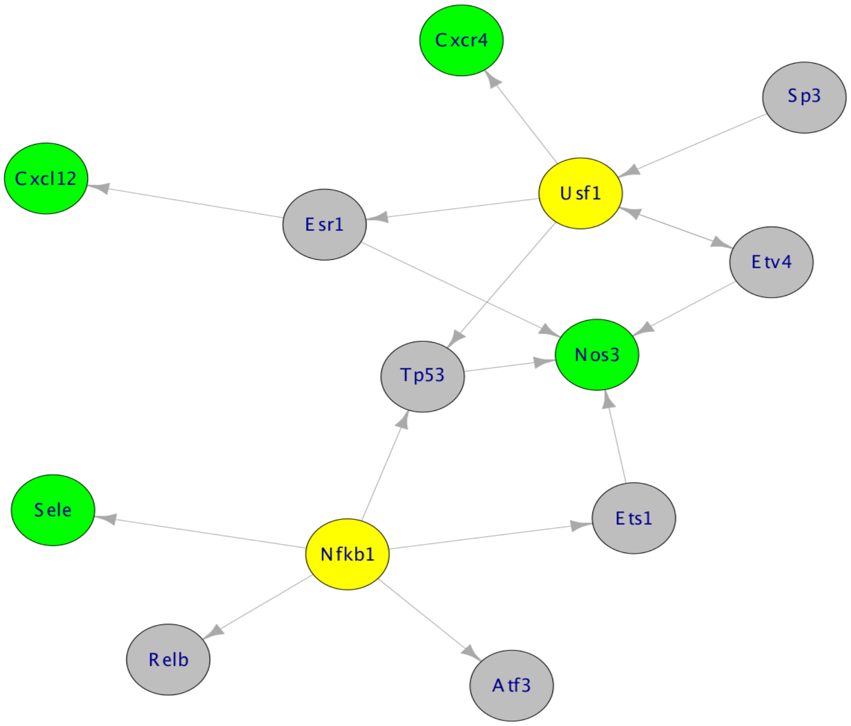


Figure 7-6 A gene network showing four EPC specific genes and their relationship to our constructed diabetic EPC GRN. The four genes: *eNOS* (*NOS3*), *SDF-1* (*CXCL12*), *CXCR4*, and *SELE* are marked in green. The two TCFs from the diabetic EPCs GRN regulating them are marked in yellow. Other TCFs regulated by the two TCFs (yellow) and also regulating those 4 genes are marked in grey.

In conclusion we were able to detect specific genes that are affected by early stages of diabetes in BM in$^-$/VEGF-R2$^+$ progenitor cells. Microarray experiments complemented by analysis methods were used in this study but no verification was performed, thus further research is warranted to confirm the results. To our knowledge, this is the first report that predicts and unravels a gene regulatory network that is specific to diabetic endothelial progenitor cells in BM. This novel GRN consists of 11 main well documented diabetic genes and 47 TCFs/genes that are regulating/regulated by those genes directly. It appears that *PIK3C2* and *PTF1A* are up regulated under diabetic conditions while *CLCNKA* is down regulated. Such changes seem to be the results of diabetic TCFs mainly *FOXO1*, *PPARα*, *PPARγ*, and *NFKB1* that controls those three diabetic genes. These findings may lead to novel therapeutic strategies for mobilization of EPCs and the treatment of diabetic vascular complications such as diabetic retinopathy, nephropathy and cardiovascular disease.

8. WGS and DNA microarray phylotyping of MRSA strains

This chapter is a shortened version of the following publication:

- Mohamed Hamed, Daniel Patrick Nitsche, Ulla Ruffing, Matthias Steglich, Janina Dordel, Duy Nguyen, Jan-Hendrik Brink, Gursharan Singh, Mathias Hermann, Ulrich Nubel, Volkhard Helms, and Lutz von Muller, Whole Genome Phylotyping and Microarray Profiling of nasal and blood stream Methicillin-Resistant Staphylococcus aureus isolates: Clues to phylogeny and invasiveness. Infection, Genetics and Evolution, 2015.

Synopsis

On the genomic mutation regards, we also presented an NGS pipeline to identify core-genome SNPs and genetic variations between two phenotypic groups in a similar analogy to somatic mutations between the healthy and disease cohorts. Since Whole Genome Sequencing data of tumor and healthy human samples were not accessible, the NGS pipeline was utilized on two groups of MRSA bacterial isolates (nasal and invasive) to investigate the phylogenetic positions of the recently emerged t504 clone (Spa-type t504) in the Saarland province of Germany and to better understand the infectivity mechanism of the invasive group as a prototype example for "from genotype to phenotype" studies.

Abstract

Hospital-associated methicillin resistant *Staphylococcus aureus* (MRSA) is frequently caused by predominant clusters of closely related isolates unresolvable by routine diagnostic typing methods. Whole genome sequencing (WGS) and DNA microarray (MA) now allow for better discrimination within a prevalent clonal complex (CC). This single center exploratory study aims to distinguish invasive and non-invasive MRSA isolates with similar genetic background into phylogenetic- and virulence-associated genotypic subgroups by WGS and MA. A cohort of twelve blood stream and fifteen nasal MRSA isolates of clonal complex 5 (CC5) (*spa*-types *t003* and *t504*) was selected.
Rooted phylotyping based on core-genome SNP WGS data revealed the regional clustering of two closely related CC5 isolate subgroups (*clade t504* and clade1 *t003*) which could be discriminated from other regional *t003* isolates and also from geographically unrelated CC5 MRSA reference isolates. However, phylogenetic subtyping was not associated with invasiveness when comparing blood stream and nasal isolates.

Clustering of MA profiles was not concordant with WGS phylotyping of CC5 MRSA isolates, but MA could discriminate subgroups of nasal and blood stream origin. Among the new putative virulence associated genes identified by WGS, the strongest association with invasiveness of blood stream infections was shown for *ebh*B gene.
Integrated analysis of core-genome in combination with accessory genome data enables in depth analysis of highly related MRSA isolates with subtyping according to phylogeny and presumable also to virulence and invasiveness *in vivo*.

8.1 Background

Approximately 20% of the healthy population is intermittently or persistently colonized with *Staphylococcus aureus* which is a well-known facultative pathogen causing localized and also generalized invasive infections [339]. Transition from colonization to invasive infection is associated with disruption of barrier and immune functions and also with the presence of bacterial virulence factors and mutations of its genetic determinants [340]. Especially, infections due to methicillin-resistant *S. aureus* (MRSA) cause a significant disease burden and also increased hospital costs [341, 342]. Prevention of MRSA infections requires early MRSA detection, decolonization, and appropriate infection control policies. Since the 1990's MRSA is further responsible for a growing number of community-acquired infections [343, 344]. Most invasive MRSA infections are related to previous colonization with the same strain [345], however, detailed knowledge of the genetic background related to transition of nasal MRSA carriage to invasive MRSA blood stream infections is still limited.

Various molecular typing methods have been developed to discriminate epidemic strains for outbreak control, epidemiological surveillance, and also for prevention of further transmission [346, 347]. Highly discriminative techniques are required for detailed analysis of local outbreaks because MRSA infections are dominated in each region by few phylogenetically related clones of the same clonal complex (CC) and sequence type (ST) [348]. Phylogeny is regularly analyzed to the CC and ST level by multilocus-sequence typing (MLST); however, this method is cumbersome and not discriminative between closely related strains. Alternatively, single locus sequence typing, *i.e. spa*-typing [349, 350], allows for discrimination of *S. aureus* isolates; however, also *spa*-typing is not very discriminative for hospital-associated infections and is not directly linked to phylogeny of isolates. The majority of MRSA isolates in German hospitals were assigned to *spa*-type *t003* isolates of CC5 [351, 352] which limits application of standard typing methods for detailed epidemic investigation and the indexing of variations for phylogenetic arrangements and population based examination [353]. Recently, the appearance of a new highly prevalent CC5 *spa*-type (*t504*) was detected apart from *spa*-type *t003* in a hospital admission prevalence screening study in the State of Saarland in Southwest Germany [354]. The structure of repeat elements was similar between local *t504* and *t003* strains; however, the phylogenetic relationship between the local *t504* and other CC5 German isolates remains still to be elucidated by more detailed broad-range phylogenetic analysis.

Compared to MLST and *spa*-typing, microarray (MA) and whole genome sequencing (WGS) dramatically enhanced discriminatory power of genotyping, and recently both technologies have become accessible also for larger scale typing purposes. For example, WGS gives valuable insight into MRSA transmission chains *e.g.* in intensive care units [355] and has already been used for the characterization of outbreaks [346, 356, 357]. It can also be implemented for detection of phenotypic properties on a genotypic base such as antibiotic resistance [358]. Although new broad-range genetic techniques may now allow for virulence assignment of clinical isolates [359], the knowledge for defined virulence-associated genotyping is still limited. MRSA virulence is caused by known and presumably also by still unknown virulence determinants and also by regulatory processes [360]. In particular, MRSA strains of the same CC may contain pathogenic patterns in a very similar genetic context, and these genotypic differences may contribute to variable virulence of invasive and non-invasive MRSA strains [361]. In line, several virulence factor (VF) online catalogs (such as the *PATRIC* [362] and *VFDB* [363] databases) were developed to affiliate information on the virulence factors in numerous organisms, species and related strains with whole genome sequence analysis of clinical isolates [364].

The goal of the present single center study was to compare nasal and blood stream MRSA isolates of the predominant CC5 by genotyping using WGS and MA. The study design included a dual approach using the core-genome single nucleotide polymorphism (SNP) WGS approach [353, 365] as well as a MA approach with a specific focus on the presence or absence of accessory gene signals. Application of WGS led to the clear discrimination of regional phylogenetic clades distinct from geographically unrelated strains of the same CC. However, invasiveness was not associated with phylogeny but with mutations of virulence factors in the core and the accessory genome.

8.2 Methods

8.2.1 MRSA CC5 isolates.

Fifteen nasal colonization isolates [366], and twelve invasive blood stream isolates from patients of the University of Saarland Medical Center (subsequently referred to as nasal [NAS] and invasive [INV] isolates, respectively) were included; WGS results were compared also to four German CC5 *t003* MRSA reference isolates provided by the

German Reference Laboratory for Staphylococcus aureus infections; all clinical MRSA isolates belonged to *spa*-types *t003* and *t504* of CC5.

8.2.2 Whole genome sequencing

Whole genome sequencing of *MRSA* DNA was performed using an Illumina MiSeq (HZI in Braunschweig, Germany) producing paired-end reads of 251 basepair lengths with an average coverage of 110-fold. In-depth bioinformatics analysis of this data was performed using the developed automated pipeline described previously in chapter 3 (Figure 3-19). Briefly, reads were mapped against the complete reference genome of *S. aureus* CC5 strain *NC_017340.1* (http://www.ncbi.nlm.nih.gov/nuccore/NC_017340) using the short read alignment version of the Burrows-wheeler Aligner (*BWA*) algorithm [175]. Both duplicate reads and reads with low mapping quality (< 30) were filtered out and the final alignments were sorted. Genetic mutations were called using the VarScan2 tool [177]. Phylogenetic analysis was restricted to the consensus sequence of the highly conserved core-genome. Therefore, variants that occurred in repetitive sequences and mobile genetic regions were masked for phylogenetic analysis.

8.2.3 Phylogeny construction

Core-genome SNPs from coding and non-coding genomic regions were extracted from the consensus sequences and a phylogenetic representative SNP matrix was generated. Subsequently, a phylogenetic tree was constructed using the Maximum Likelihood method as implemented in SeaView [178] and rooted using the CC5 ST5 reference genome N315 (http://www.ncbi.nlm.nih.gov/nuccore/NC_002745). Trees were displayed and annotated using FigTree (http://tree.bio.ed.ac.uk/software/figtree/). For comparison, four representative German CC5 isolates were also included for direct comparison with the local strains and in order to enrich the phylogenetic tree. The dendrograms of the DNA microarray were produced using the hierarchical clustering algorithm using average linkage and Euclidian distance in the R suite [173].

8.2.4 Genetic variations between invasive and nasal samples

The significance of the genetic variations between each isolate pair (invasive and nasal) was evaluated by VarScan on the basis of the sequence reads through Fisher's exact test using a significance level or p-value threshold of 0.05. Successfully passed variants were collected and annotated to the corresponding genes in the reference genome. Subsequently, the variants were grouped by position, and the occurrence of each variant was noted. The identities of 548 known virulence-related genes were derived from the virulence factor (VF) databases *NIAID* Pathogen Annotation Browser [367], *PATRIC [362]* , and *VFDB* [363]. Over representation analysis for enrichment with functional GO terms, KEGG pathways, and INTERPRO protein families was performed using the DAVID tool [367].

8.2.5 DNA microarray analysis

DNA extraction (Qiagen, Hilden, Germany), and microarray analysis using IdentiBAC MA (Alere, Jena, Germany) was performed according to the manufacturer´s instruction as previously described [366]. Data processing and bioinformatics grouping according to similarity of genetic profiles was done accordingly.

8.3 Results

Whole genome sequencing produced paired-end reads of 251 bp length at about 110-fold coverage. 81-87% of the sample reads could be mapped against the reference CC5 genome *NC_017340.1* (http://www.ncbi.nlm.nih.gov/nuccore/NC_017340) (spa-type *t003*) with a maximal possible mapping error of 0.1 %. By masking repetitive sequences and mobile genetic regions, the first part of our analysis focused on the so-called core-genome region and included reconstruction of the strain phylogeny and SNP analysis.

8.3.1 Phylogenetic analysis

Phylogenetic analysis based on the 1112 core-genome SNPs of WGS data showed that all isolates with *spa*-type *t504* clustered separately from type *t003* forming a single clade (*Clade t504*) (Figure 8-1). The phylogenetic distribution of *t003* was generally more diverse but cluster analysis revealed a distinct clade of nine local *t003* isolates (*Clade1 t003*) without direct connection to geographical unrelated German reference isolates. Also, the remaining eight local isolates without particular clustering were distributed without direct link to the geographical unrelated German reference isolates (*Other t003*). This confirms high diversity in phylogenetic arrangements of the *t003* strain in geographically distinct regions [352, 353].

8.3.2 SNP analysis

SNP calling of the WGS data identified genetic variations in 535 unique genomic positions outside of mobile genetic elements and repetitive sequences between all pairs of the 12 invasive isolates and the 15 nasal ones. These 535 positions include 479 SNPs and 56 Indels. Clade *t504* (36 ± 7 mutations) and clade1 *t003* (43 ± 8) isolates contained fewer mutations than other regional *t003* isolates (56 ± 11) (Figure 8-2). However, only 40% of genetic mutations occurred in annotated regions and involved 176 genes. Among these genes, 18 genes containing 24 variants were previously characterized as virulence–related genes in the VF catalogs (*tcaA, rnr, saeS, sasA, msrR, ssaA, capA, arlS, hlgB, sdrD, aur, hysA, isdE, isdF, hlb, essA, atl, and lip*). Interestingly, all of these 18 known virulence-related genes had variants in at least one invasive sample; yet, no such variants were recorded in the nasal isolates with according genes. Twenty genes showed variants in at least two invasive isolates, yet again, in genes from strains of nasal origin these variations were absent. In the following, we will refer to such genes as ‘twice mutated genes’. Such twice mutated genes of invasive isolates include two known virulence-related genes (*atl, hlb)* and 18 genes that have not been associated with virulence so far (*ebhB, pfoS/R, glpF, feoB, yvcP, sbnD, mutS2, prkC, miaA, thrC, trpD, gnd, sodA, tagG, kdpD, metT, tcaB, opp-1F*) (Supplementary Table E-2).

Yet, twice mutated gene variants of invasive versus nasal strains failed to reach statistical significance (*P-values* of 0.20 and 0.07, Fisher’s exact test); only the gene *ebhB* (coding for a putative Staphylococcal surface anchored giant protein) showed genetic variations at 7 positions (1482083, 1484403, 1487821, 1491377, 1499271, 1502542, 1503065) thus covering most of the entire gene. Each of these mutations occurred in at least 2 out of 12 invasive strains, but not in nasal strain. The difference of *ebhB* mutations between invasive (7 out of 12 invasive strains) and nasal isolates (0 out of 15 nasal strains) was statistically highly significant (Fisher´s exact test, *P-value* = 0.0009).

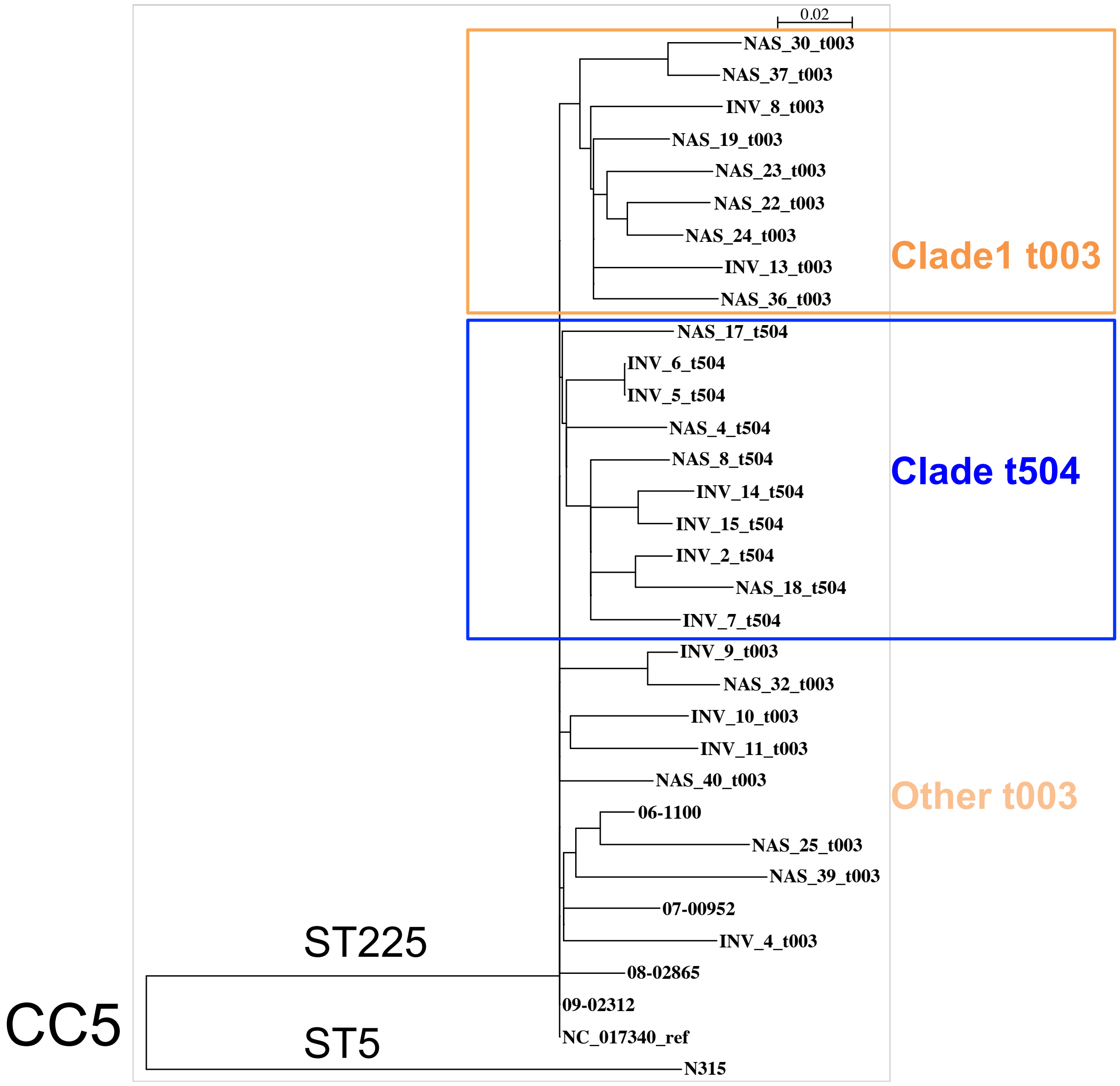
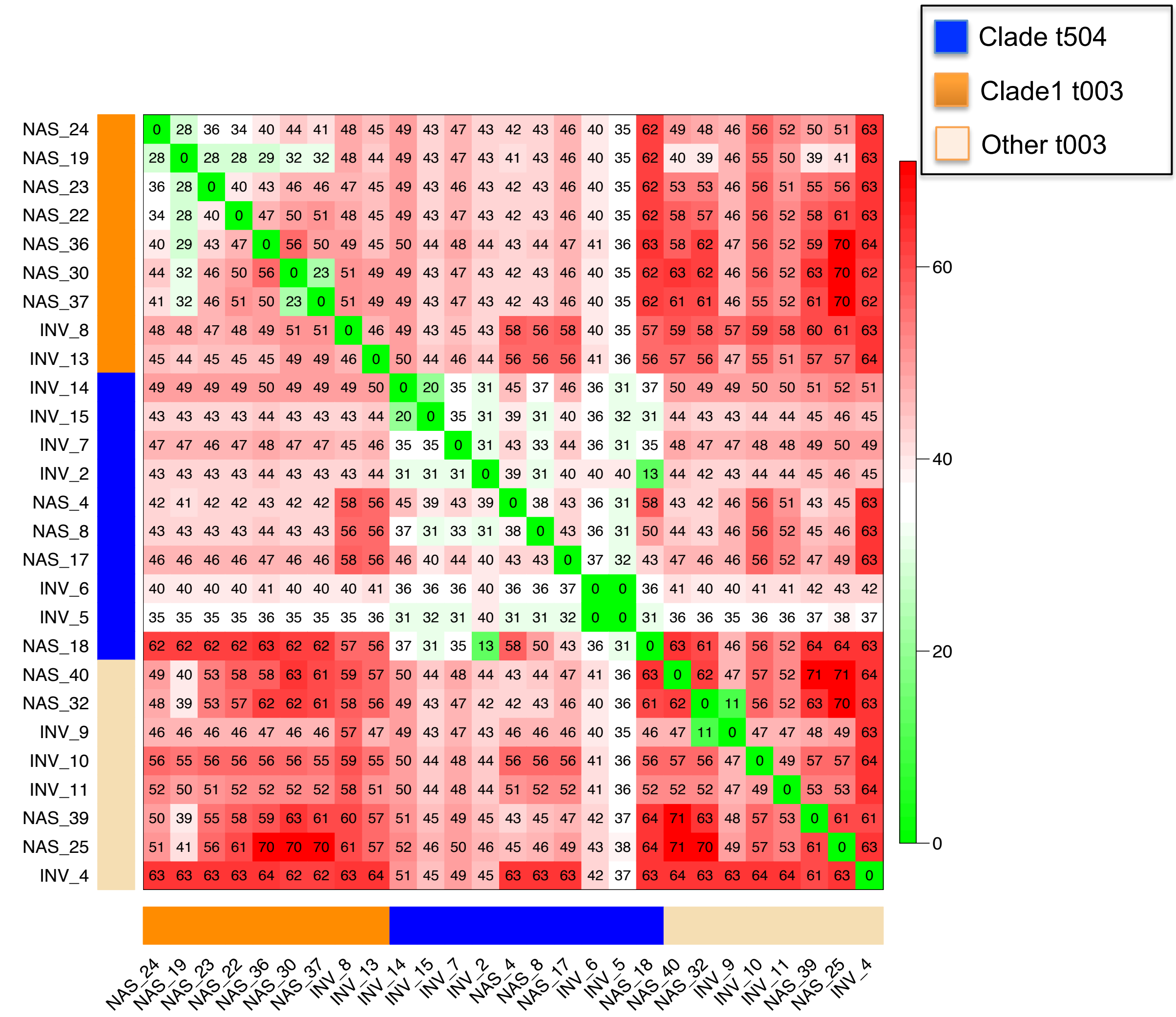


Figure 8-1 Phylogenetic tree of 27 invasive and nasal S. aureus CC5 strains collected in Saarland as well as four representative S. aureus reference isolates in Germany based on the core-genome SNP approach. The tree was rooted with the genome sequence from isolate N315 (ST5; NCBI accession no. NC_002745).

To get further insight in the genomic location of the twice mutated genes, we analyzed their genomic distance to known virulence-associated genes by use of a Manhattan plot (Figure 8-3). Mutations in genes of candidate virulence variants (twice mutated genes marked in green) were significantly closer to variants in 18 known virulence genes (marked in red) than to random *SNP* positions (*P-value* = 0.035).

The 18 at least twice mutated genes were related to metabolic pathways and functional biological process terms of the Gene Ontology and tested their affiliation to protein families using statistical term over representation analysis. The gene products of *sbnD* and *tcaB* belong to protein family tetracycline resistance protein, *TetA*/multidrug resistance protein *MdtG* (*INTERPRO: IPR001958*) that prevents the antibiotic tetracycline from inhibiting bacterial protein synthesis. The genes *trpD* and *kdpD* take part in the *KEGG* pathway *two component system.* However, the majority of these genes are not characterized so far in the annotation databases (Supplementary Table E-1).



	NAS_24	NAS_19	NAS_23	NAS_22	NAS_36	NAS_30	NAS_37	INV_8	INV_13	INV_14	INV_15	INV_7	INV_2	NAS_4	NAS_8	NAS_17	INV_6	INV_5	NAS_18	NAS_40	NAS_32	INV_9	INV_10	INV_11	NAS_39	NAS_25	INV_4
NAS_24	0	28	36	34	40	44	41	48	45	49	43	47	43	42	43	46	40	35	62	49	48	46	56	52	50	51	63
NAS_19	28	0	28	28	29	32	32	48	44	49	43	47	43	41	43	46	40	35	62	40	39	46	55	50	39	41	63
NAS_23	36	28	0	40	43	46	46	47	45	49	43	46	43	42	43	46	40	35	62	53	53	46	56	51	55	56	63
NAS_22	34	28	40	0	47	50	51	48	45	49	43	47	43	42	43	46	40	35	62	58	57	46	56	52	58	61	63
NAS_36	40	29	43	47	0	56	50	49	45	50	44	48	44	43	44	47	41	36	63	58	62	47	56	52	59	70	64
NAS_30	44	32	46	50	56	0	23	51	49	49	43	47	43	42	43	46	40	35	62	63	62	46	56	52	63	70	62
NAS_37	41	32	46	51	50	23	0	51	49	49	43	47	43	42	43	46	40	35	62	61	61	46	55	52	61	70	62
INV_8	48	48	47	48	49	51	51	0	46	49	43	45	43	58	56	58	40	35	57	59	58	57	59	58	60	61	63
INV_13	45	44	45	45	45	49	49	46	0	50	44	46	44	56	56	56	41	36	56	57	56	47	55	51	57	57	64
INV_14	49	49	49	49	50	49	49	49	50	0	20	35	31	45	37	46	36	31	37	50	49	49	50	50	51	52	51
INV_15	43	43	43	43	44	43	43	43	44	20	0	35	31	39	31	40	36	32	31	44	43	43	44	44	45	46	45
INV_7	47	47	46	47	48	47	47	45	46	35	35	0	31	43	33	44	36	31	35	48	47	47	48	48	49	50	49
INV_2	43	43	43	43	44	43	43	43	44	31	31	31	0	39	31	40	40	40	13	44	42	43	44	44	45	46	45
NAS_4	42	41	42	42	43	42	42	58	56	45	39	43	39	0	38	43	36	31	58	43	42	46	56	51	43	45	63
NAS_8	43	43	43	43	44	43	43	56	56	37	31	33	31	38	0	43	36	31	50	44	43	46	56	52	45	46	63
NAS_17	46	46	46	46	47	46	46	58	56	46	40	44	40	43	43	0	37	32	43	47	46	46	56	52	47	49	63
INV_6	40	40	40	40	41	40	40	40	41	36	36	36	40	36	36	37	0	0	36	41	40	40	41	41	42	43	42
INV_5	35	35	35	35	36	35	35	35	36	31	32	31	40	31	31	32	0	0	31	36	36	35	36	36	37	38	37
NAS_18	62	62	62	62	63	62	62	57	56	37	31	35	13	58	50	43	36	31	0	63	61	46	56	52	64	64	63
NAS_40	49	40	53	58	58	63	61	59	57	50	44	48	44	43	44	47	41	36	63	0	62	47	57	52	71	71	64
NAS_32	48	39	53	57	62	62	61	58	56	49	43	47	42	42	43	46	40	36	61	62	0	11	56	52	63	70	63
INV_9	46	46	46	46	47	46	46	57	47	49	43	47	43	46	46	46	40	35	46	47	11	0	47	47	48	49	63
INV_10	56	55	56	56	56	56	55	59	55	50	44	48	44	56	56	56	41	36	56	57	56	47	0	49	57	57	64
INV_11	52	50	51	52	52	52	52	58	51	50	44	48	44	51	52	52	41	36	52	52	52	47	49	0	53	53	64
NAS_39	50	39	55	58	59	63	61	60	57	51	45	49	45	43	45	47	42	37	64	71	63	48	57	53	0	61	61
NAS_25	51	41	56	61	70	70	70	61	57	52	46	50	46	45	46	49	43	38	64	71	70	49	57	53	61	0	63
INV_4	63	63	63	63	64	62	62	63	64	51	45	49	45	63	63	63	42	37	63	64	63	63	64	64	61	63	0

Figure 8-2 Heatmap showing the number of genetic variations between each pair of isolates.
The highest similarity was found for clade t504 followed by t003 clade I. A direct comparison of groups revealed higher similarity between t003 clade I and t504 as compared to t003 clade I and other t003 isolates.

8.3.3 Clustering based on DNA microarray

Hierarchical clustering analysis of the entire set of 330 MA probes yielded five major clusters with at least three isolates (Figure 8-4a). Among them, cluster A1 contained only invasive blood stream isolates (*P-value* =0.01, Fisher exact test) and was characterized by positive signals of *hsdSx.CC15* (allelic variant of type I site-specific deoxyribonuclease subunit), and *Q2YUB3* (unspecific efflux transporter). Cluster A2 (*P-value*=0.23) contained only *ccrB.4* positive (cassette chromosome recombinase) nasal isolates whereas clusters A3 to A5 were without clear predisposition to invasive *vs.* nasal isolates. When grouping was restricted to virulence genes annotated in the virulence factor (VF) catalogues (174 genetic probes), six clusters were identified (Figure 8-4, B1-6). *B3* (*P-value* =0.08) is characterized by positive hybridization signals of *ssl01.set6_probe2_11* and *ssl01.set6..MRSA252* (allelic variants of the staphylococcal superantigen-like protein 1 termed *set6*) and encompasses only invasive strains. *B2* (*P-value* =0.66) and *B4* (*P-value* =0.11) contain only nasal strains, whereas the remaining clusters represent both invasive and nasal isolates. For comparison, read counts for these genes were related to those found in the WGS data.

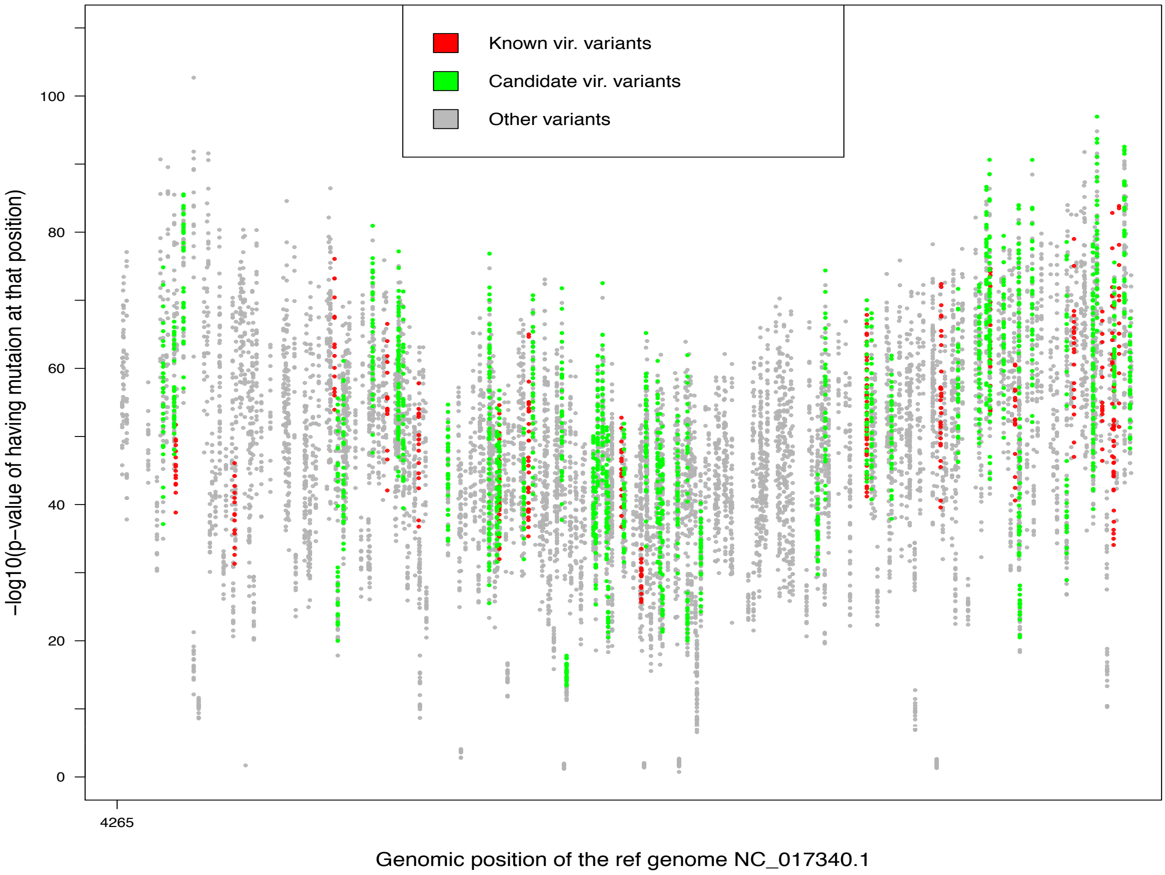


Figure 8-3 Manhattan Plot showing the genomic distribution of the 535 variants between each pair of isolates. Related genes of 24 variants occurring in known virulence are marked in red. 54 variants that occur exclusively in at least two virulent isolates are marked in green. By testing against randomized SNP positions, we showed that mutated still undefined genes (green) are significantly closer to known virulence genes (red) than expected by chance (P-value = 0.035).

Two out of these four genes (*hsdSx.CC15,* and *Q2YUB3*) were not annotated in the reference genome. Gene *set6* has 10 allelic variants; thus, a simple comparison of read numbers across isolates was not helpful. Gene *ccrB* had on average an about two-fold coverage in strains NAS22-NAS24 belonging to cluster A2 (maximal read coverage is about 200, average read coverage is 176) compared to the strains INV4, INV6, and INV8 forming cluster B3 (maximal read coverage is about 120, average read coverage is 93). This result matches the exclusively positive signal for this gene in strains of cluster A2 based on the MA experiment. Next, the interrelation between the phylogenetic clades, dendrogram groups and mutations occurring in known virulence genes and at least twice-mutated genes was addressed.

Table 8-1 reveals that the different phylogenetic clades are associated with mutations occurring in known virulence genes as well as in twice-mutated genes but not with the positive or negative hybridization signals for any of the DNA microarray genes.

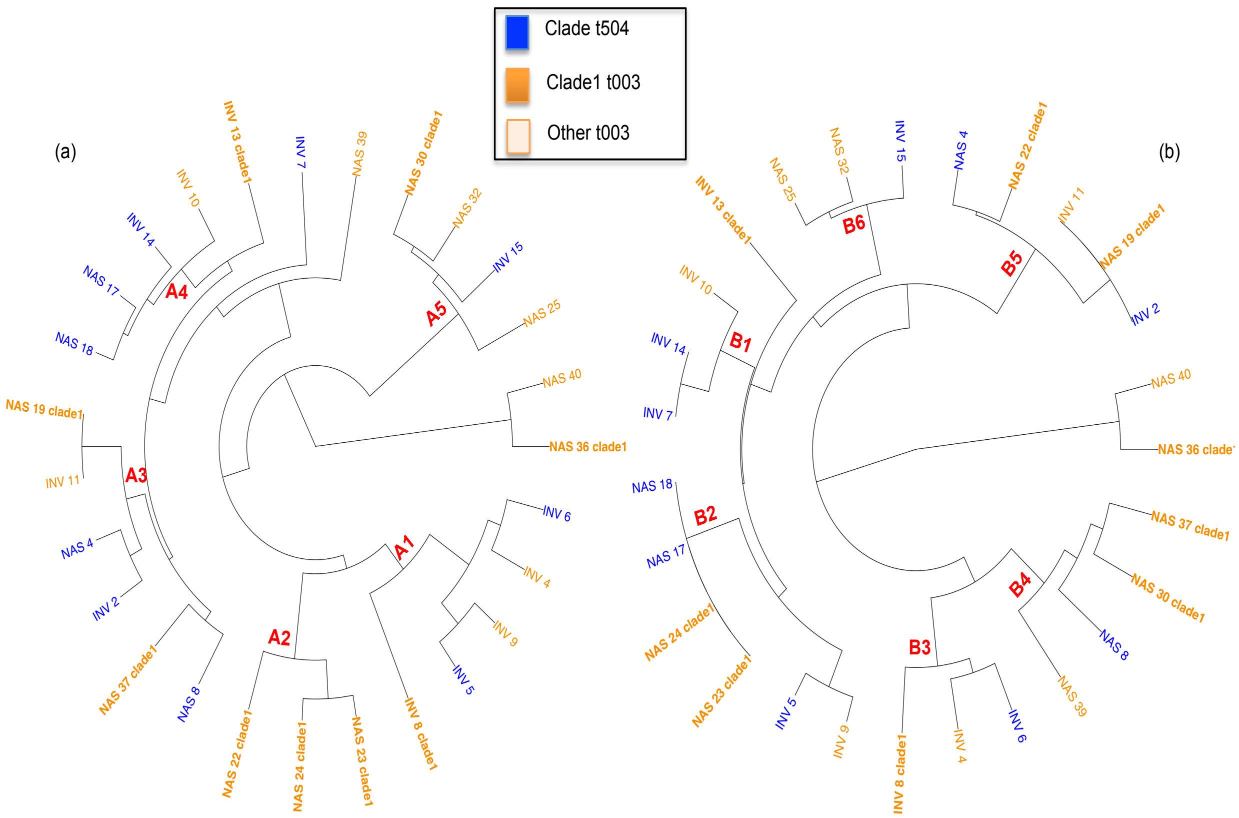


Figure 8-4 Dendrogram based on hierarchical clustering of DNA microarray (MA).
Invasive (INV) and nasal (NAS) isolates were indicated and also strain assignment by spa-typing and core-genome SNP analysis into phylogenetic clades was indicated. (a) Genetic profiles of all MA genes/alleles were used (n = 330) while (b) alternative analysis was restricted to 174 probes of annotated genes in the virulence catalogs.

8.4 Discussion

Here we confirm that WGS with core-genome SNP analysis is applicable for analyzing the evolutionary distance between closely related CC5 MRSA strains. Using less discriminatory methods the provenience of the recently evolved regional *spa*-type *t504* strains remained elusive [354, 366]. Based on WGS we can now prove that the *t504* strains were of clonal origin with common ancestors to the highly abundant *t003* group. Interestingly, a second clade of CC5 isolates was identified by WGS without direct association to other *t003* strains including German reference strains of other provenience. Therefore we hypothesize that – similarly to the *t504* clade – a second clade of CC5 isolates evolved in the region of Saarland (*t003* clade 1) which has not been identified before by less discriminatory typing methods. In the present study regional clades (*t504* and *t003* clade 1) were detected in the population during a hospital admission study [354] and not following MRSA transmission in the same hospital [355]. The appearance of regional clades argues for MRSA spreading in regional health-care associated networks. The phylogenetic tree revealed that the *t003* strains were more diverse than the *t504* strains. However, the close distance between *clade t504* and *clade t003* suggests that *t504* strains might have evolved from common ancestors. CC5 isolates from other regions of Germany were highly distant according to cluster analysis showing high phylogenetic diversity.

Although WGS is able to determine genome sequences at fast pace and affordable costs it has argued that bioinformatics techniques are still lacking that enable to extract information about the true virulence potential of an organism from sequence data alone [368]. As a possible means to detect functional relevant mutations, they proposed setting up a virulence-assaying framework based on medium-throughput phenotypic characterization of candidate strains, *e.g.* by assaying their adhesion to fibronectin or cytolytic activity. Applying WGS to RN4220 strains could identify a number of SNPs affecting the general fitness of the bacteria [369]. Most relevant to the present work were the recent studies based on WGS who were able to associate the toxicity of 90 *MRSA* ST239 isolates with around 100 statistically significant SNPs [359] and studies investigating the evolutionary dynamics of *S. aureus* during long-term carriage and transition from nasal carriage to invasive infection [340, 360, 370, 371].

Table 8-1 Association of phylogenetic clades to the known virulence factor genes and the twice mutated genes (that have variants in at least two invasive isolates but none in isolates of nasal origin)

	Mutated genes detected by WGS	
Phylogenetic groups (WGS)	**Known virulence genes**	**Twice mutated genes**
Clade1 t003	sdrD, msrR, hysA, tcaA, ssaA, sasA	
Clade t504	essA, saeS, atl, isdF, hlb, lip	sbnD, mutS2, prkC, glpF, miaA, thrC, trpD, ebhB, sodA,pfoS/R, tagG, kdpD, metT, tcaB, opp-1F, yvcP
Other t003	capA, rnr, isdE, arlS, hlb, hlgB, aur, sasA	gnd, feoB

Here, we followed a similar strategy that is based on identifying genetic variations occurring exclusively either in the group of invasive blood stream or in the group of colonizing nasal isolates. We identified the genetic variations between all pairs of blood stream and nasal isolates and we demonstrated that invasiveness was not associated with phylotyping (core-genome SNP analysis) but with characteristic mutations of known and presumed new candidate virulence genes.

We identified SNPs between blood stream and nasal isolates (twice mutated genes), but for most mutated genes the difference was not significant presumably due to the limited number of isolates in present exploratory study. However, a significant association with invasiveness was detected for mutations in the giant protein *ebh*B gene. The *ebh* gene is the largest open reading frame on the *S. aureus* (33kb) encoding for the Giant Staphylococcal Surface Protein (GSSP) which is a membrane anchored protein capable for binding of matrix components, to protect the cell against osmotic pressure changes and to control agglutination [370, 372]. Analysis of *ebhA* and *ebhB* in some genomes (Mu50, N315) showed that the original single open reading frame was disrupted by a frameshift mutation leading to their permanent separation [373].

The genetic vicinity between known virulence factors and twice mutated genes of unknown pathogenicity could be interpreted as a sign of functional relatedness in respect to pathogenicity; however, this hypothesis remains to be confirmed by experimental studies. Mutations in pathogenicity-associated genes may become selected due to enhanced infectivity, increased virulence [374] or increased fitness and better adaptation processes. Routine comparison of WGS data on blood stream and nasal CC5 MRSA was limited in the present study because SNP analysis was restricted to mutations of the core genome due to technical reasons. It has been pointed out that mobile genetic elements encoding for antibiotic resistance and virulence [372] may be responsible for dynamic phenotypic changes. Hence, they were not covered in the present SNP analysis focusing on virulence-associated mutations in the core-genome [370].

In contrast to WGS with core-genome SNP approach, clustering of MA data did not allow sub-clustering according to phylogenetic distances (Figure 4a-b). Instead, cluster analysis based on MA apparently identified distinct clusters of invasive isolates according to the presence of characteristic microarray hybridization profiles in accessory genes outside the core-genome. However, discrimination of most MA based subgroups was not significant presumably due to low number of isolates in the present exploratory study. The presence of known and presumed virulence genes in related and unrelated MRSA CC5 subgroups strengthens the hypothesis that MRSA virulence and invasion was not associated with phylotyping but with characteristic mutations of genome[359].

In conclusion, we have shown that whole-genome sequencing analyzed by phylogeny-oriented mapping and SNP-analysis of the core-genome as well as DNA-hybridization data detected by microarrays are able to provide important insight of complementary nature into evolution and virulence of MRSA CC5. This approach identifies potential new candidate virulence genes requiring confirmation by independent experimental studies. We also demonstrated that increased virulence and invasiveness was not associated with phylogeny. Coupled core- and accessory genome WGS analyses require additional tools for better discrimination of infection associated and commensal MRSA strains.

9. Conclusion and outlook

Synopsis

In this chapter, we summarize the results achieved in this thesis and discuss the current limitations of the introduced approaches and directions for further improvements. Moreover, we shed light on possible implications for future research and follow-up work for this thesis.

9.1 Accomplished work

The enormously increasing availability of transcriptomic and epigenomic data from different biological experiments allow for deep and comprehensive integrative analysis. Functions of the molecular elements (genes, proteins, mutations, miRNAs, siRNAs,..etc) that represent the entities of such genomic data are highly connected with the underlying cellular malfunctions and disease pathways. Moreover, these molecular elements interact with each other forming a complex interwoven regulatory machinery that govern the cellular pathways of biological functions or the pathology of diseases. Therefore, revealing these critical molecular interactions in complex living systems is being considered as one of the major goals of the systems biology revolution nowadays.

In this dissertation, we have introduced practical computational approaches implemented as freely available software tools to integrate heterogeneous sources of large-scale genomic data and unravel the combinatorial regulatory interactions between different molecular elements. These proposed approaches were applied to investigate the molecular mechanisms of cellular differentiation as an example for biological processes and human breast cancer, and diabetes as examples for complex diseases.

First, the automated GRN pipeline constructs the genomic regulatory machinery of a cell from expression, sequencing, and annotation datasets through three modules (Chapter 3). The GRN pipeline starts with building the weighted co-expression network, then searches for TFs motifs in the sequences of the network genes as well as harvests the known interactions whose source and end nodes are involved in the co-expression networks. Next, the confirmed interactions between the two previous steps are subjected as a prior network to a Bayesian learner module. The selection of the Bayesian approach as a reconstruction method was based on the question of how to infer interactions while making use of what is already known. Bayesian network learning algorithms allow using initial network as a prior knowledge to guide the learning process. Moreover, Bayesian networks enable dealing with noises that are inherent in microarray data and to model hidden nodes in the network. Application of the GRN pipeline on gene expression and sequence data of blood cell differentiation (hematopoiesis, chapter 5), mouse diabetes samples (chapter 7), and human breast cancer data (chapter 6) demonstrated its usefulness in unraveling the architecture and features of the corresponding GRN network. We have also assessed the performance of the GRN pipeline by benchmarking its results against two other statistical methods (chapter 6) and found plausible overlaps that confirm the efficacy of the Bayesian approach in learning the GRN topology. Nevertheless, we refer to important limitations of Bayesian learning algorithms in the following section.

The three modules of the GRN pipeline are implemented as separated software components (plugins) and hosted by our software framework Mebitoo for workflow automation. We note that coupling the third module is still in progress. Mebitoo is a software application suite written in Java that is based on the Netbeans Rich-Client platform (RCP) project that can easily be extended with additional functionality deployed as modules. Since the Mebitoo framework implements a uniform plugin interface, automated data processing can be invoked using a task execution interface in order to queue multiple operations of different modules and process datasets in parallel.

Mebitoo is appropriate for inexperienced users, researchers without programming knowledge as well as scientific programmers, and developers. For the first group, we provide an easy and friendly GUI that guides the user to sequentially define his tasks (every task represents one-time running module) and gets the final results in one–click press button. For advanced users with knowledge in java programming, Mebitoo can be used as a ready hosting workflow automation framework for coupling more new bioinformatics add-on plugins or modules.

While the capabilities of the GRN pipeline are limited to capture only gene interaction information at the transcriptional level using gene expression and gene sequencing data, we further extended it to a general integrative network-based approach that involves also post-transcriptional interactions and reports the computational analysis of gene and miRNA transcriptomes, DNA methylome, and somatic mutations. This aimed at identifying putative disease drivers and novel targets for therapeutic treatment. This approach has been applied to breast cancer data and was able to reveal the strong association between regulatory elements from four different genomic data sources. This integrated molecular analysis enabled by this approach substantially expands our knowledge base of prospective genomic drivers of genes, miRNAs, and mutations highlights candidates for further investigation in the wet lab as novel targets for breast cancer treatment (Chapter 6). Also by benchmarking the provided approach, it can be applied in a similar fashion to other cancer types, complex diseases, or for studying cellular functions where such multi-dimensional datasets are available.

Regarding to the incorporation of somatic mutations with other genomic data sets, a stand-alone pipeline named “SnvDMiR” was implemented to explore possible genomic proximity relationships between somatic variants and both differentially methylated CpG sites as well as differentially expressed miRNAs. Further analysis on somatic variants that occur in close genomic vicinity to the deregulated miRNAs or CpG sites revealed mutations that are candidate drivers of oncogenic processes in breast cancer. With respect to the genomic mutations, we also presented an NGS pipeline to identify core-genome SNPs and genetic variations between two groups in a similar analogy to somatic mutations between the healthy and disease cohorts. Since Whole Genome Sequencing data of tumor and healthy human samples were not accessible, the NGS pipeline was utilized on two groups of *MRSA* bacterial isolates (nasal and invasive) to investigate the phylogenetic positions of the recently emerged t504 clone (Spa-type t504) in the Saarland province of Germany and to better understand the infectivity mechanism of the invasive group as an example of a “from genotype to phenotype” study (Chapter 8).

Motivated by this, we developed TFmiR as a freely available web server for deep and integrative downstream analysis of combinatorial regulatory interactions between TFs/genes and miRNAs that are involved in human disease pathogenesis. TFmiR helps to better elucidate disease cellular mechanisms on the molecular level from a network perspective. The TFmiR web server is based on user-provided sets of deregulated genes and/or miRNAs regardless of the data producing technologies of either microarray experiments, NGS, or PCR. The usefulness of TFmiR was confirmed by constructing the breast cancer–specific network and identifying literature-confirmed core regulators as well as novel hub nodes of TFs/miRNAs that could be further experimentally investigated as new potential drug targets. TFmiR was also able to characterize important TF–miRNA co-regulatory motifs whose co-regulated genes form cooperative functional modules in breast cancerogenesis. Our web server showed advances over other related web tools in terms of the extended downstream analysis, the variety of user parameters, user call scenarios, and in terms of incorporating information from various well-established regulatory databases.

In summary, the work presented in this thesis has led to the development of interesting computational approaches that are introduced to the scientific literature in non-commercial software toolkits. The provided topological and functional analyses of our frameworks as validated on cellular differentiation and complex diseases promotes our frameworks as reliable systems biology tools for researchers across the life science communities.

9.2 Limitations of this work

The GRN pipeline utilizing a Bayesian learner showed a remarkable potential to infer the network topology. However, there are some inherent concerns that need to be mentioned. For instance, the Bayesian approach doesn't allow cycles or loops in the inferred networks, whereas most genes have negative feedback effect on their own expression. Furthermore, Bayesian inference algorithms require sophisticated preprocessing procedures for the expression data such as normalization, data denoising, missing value imputation, and discretization. We didn't encounter such a concern in any of our datasets because the data used were carefully preprocessed. Another major problem of Bayesian networks is the computational difficulty and the costly processing due to the dimensionality problem of the input microarray data. The Bayesian learning algorithms are not generally suited for inferring larger networks with hundreds or thousands of genes. However, assuming a sparse nature of the GRN where each gene is regulated by relatively few genes, data partitioning methods such as clustering and biclustering techniques have been introduced to group genes into functionally homogenous clusters before applying the Bayesian learner.

As with any prediction method, the inferred interactions require experimental verification such as the standard laboratory knock-out experiments, enforced expression of the TF and monitoring the expression pattern of the target gene, and Chipseq experiment to examine whether the binding motifs of the TF are close to the TSS of the target gene. However, this was not established parallel to this work due to lack of resources and time. The validation in this work was mainly based on literature-confirmed evidences, other computational and statistical methods such as statistical significance, functional similarity, reporting phenotypes due to gene knock-outs from MGI database, and based on benchmarking and assessing the performance against similar approaches. Unfortunately, gold standard networks for hematopoiesis or human cancer were not available to systematically verify the constructed interactions using AUC ROC and precision and recall curves.

As discussed in Chapter 6, the samples used as input for the integrative network based approach are the matched and common ones between the four TCGA data sets (mRNA

Expression, DNA Methylation, miRNA expression, and somatic mutations). This means that data sets have to be consistent and belong to the same samples (cell lines/patients). Otherwise, technical variations between different samples cultured in various labs could lead to inappropriate results. A serious problem of this approach is the lack of consistent data as was reviewed in the blood cell differentiation (chapter 5). Apparently, to date not many genomic data repositories offer such consistent data for developmental cell lines or diseases. Up to our knowledge, only the TCGA portal provides consistent data for normal and tumor samples for various cancer types in human.

9.3 Outlook

Although heterogeneous sources of genomic datasets have been incorporated in this work to reverse engineer the complex regulatory networks, it is still not persuasively sufficient to build realistic dynamic models from the acquired GRN networks. This is due to the role of other cellular mechanisms, which are believed to contribute to the regulatory machinery in the cell such as epigenetic mechanisms (histone modifications, siRNA interference, regulated degradation of mRNA) and post-translational events (protein phosphorylation, processing, or localization). Hence, there is abundant room for further research on deciphering the regulatory roles of these cellular activities by incorporating representing data sets to those introduced in this work. Along the same line, the biophysics nature of the cell like the roughness characteristics of the surface markers attached to the hematopoietic stem cells seemed to influence the differentiation competency of stem cells. This opens up new avenues for future research areas on assembling these biophysics properties into the full picture of the regulation machinery. However, more deep research on the regulatory role of biophysics characteristics of the cell needs to be undertaken in advance.

Once the roles of cellular mechanisms affecting the regulatory machinery are encoded in the acquired network, it would be interesting to stimulate the network and study the network dynamics to identify master regulatory elements that are responsible for most genetic diseases and accordingly could serve as a commencing point for therapeutic treatment. Similarly in cellular programming, this would help in identifying key driver molecules and their interactions, which determine the conditions at which the cell switches to the next cell stage.

On the other hand, it is fairly acceptable within the research community to classify biological networks into gene regulatory networks, signal transduction network, metabolic networks, and protein-protein interaction network. Though, to what extent the transcriptional regulatory network can be decoupled from the other networks for the sake of reducing the complexity of biological systems. Further research work needs to examine more closely the links between the different biological networks.

Appendix A: Supplementary of Chapter 4

Table A-1: Imprinted Gene list. The last column indicates whether the maternal (M) or paternal (P) allele is expressed. P/M means that the gene exhibits species or isoform-specific patterns of imprinting: human COPG2 was reported to be paternally expressed, while this gene is maternally expressed in the mouse. Human ZIM2 is paternally expressed, whereas the murine Zim2 gene is active on the maternal chromosome. GRB10 encodes maternally, and paternally expressed isoforms. "?" in the imprinting column indicates genes for which the imprinting status is not known.

Expressed Allele	**Imprinting**		**Description**	**Gene Name**
	Mouse	**Human**		**Human (Mouse)**
M	Y	?	ankyrin repeat and SOCS box-containing 4	ASB4 (Asb4)
M	Y	?	achaete-scute complex homolog 2 (Drosophila)	ASCL2 (Ascl2)
M	?	Y	ATPase, class V, type 10A	ATP10A (Atp10a)
P	Y	?	brain-enriched guanylate kinase-associated homolog (rat)	BEGAIN (Begain)
M	Y	Y	bladder cancer associated protein	Blcap
?	(no ortholog)	Y	chromosome 15 open reading frame 2	C15ORF2
M	Y	?	calcitonin receptor	CALCR (Calcr)
M	Y	Y	cyclin-dependent kinase inhibitor 1C (p57, Kip2)	CDKN1C (Cdkn1c)
M	Y	N	copper metabolism (Murr1) domain containing 1	COMMD1 (Commd1)
P/M	Y	?	coatomer protein complex, subunit gamma 2	COPG2 (Copg2)
M	?	Y	carboxypeptidase A4	CPA4 (Cpa4)
P	Y	?	deiodinase, iodothyronine, type III	DIO3 (Dio3)
P	?	Y	discs, large (Drosophila) homolog-associated protein 2	DLGAP2 (Dlgap2)
P	Y	Y	delta-like 1 homolog (Drosophila)	DLK1 (Dlk1)
M	?	Y	distal-less homeobox 5	DLX5 (Dlx5)
M	Y	Y	GNAS complex locus	GNAS (Gnas)
P/M	Y	Y	growth factor receptor-bound protein 10	GRB10 (Grb10)
M	Y	?	histocompatibility 13	H13
M	Y	?	5-hydroxytryptamine (serotonin) receptor 2A	HTR2A (Htr2a)
P	Y	Y	insulin-like growth factor 2	IGF2 (Igf2)
M	Y	N	insulin-like growth factor 2 receptor	IGF2R (Igf2r)
P	Y	no ortholog	Impact homolog (mouse)	IMPACT (Impact)
P	Y	?	inositol polyphosphate-5-phosphatase F	INPP5F (Inpp5f)
P	Y	Y	insulin 2	INS (Ins2)
M	Y	Y	potassium voltage-gated channel, KQT-like subfamily, member 1	KCNQ1(Kcnq1)
M	Y	Y	Kruppel-like factor 14	KLF14 (Klf14)
M	Y	Y	potassium channel, subfamily K, member 9	KCNK9 (Kcnk9)
P	N	Y	lethal(3)malignant brain tumor-like protein-like	L3MBTL (L3mbtl)
P	N	Y	leucine rich repeat transmembrane neuronal 1	LRRTM1 (Lrrtm1)
P	Y	Y	MAGE-like 2	MAGEL2 (Magel2)
P	Y	Y	malignant T cell amplified sequence 2	MCTS2 (Mcts2)
P	Y	Y	mesoderm specific transcript homolog (mouse)	MEST (Mest, Peg1)
P	Y	Y	makorin ring finger protein 3	MKRN3 (Mkrn3)
P	Y	Y	nucleosome assembly protein 1-like 5	NAP1L5 (Nap1l5)
P	Y	Y	necdin homolog (mouse)	NDN (Ndn)
P	Y	Y	neuronatin	NNAT (Nnat)

P	Y	Y	paternally expressed 3; PEG3 antisense RNA (non-protein coding); zinc finger, imprinted 2	PEG3 (Peg3)
P	Y	Y	paternally expressed 10	PEG10 (Peg10)
P	Y	no ortholog	paternally expressed 12	(Peg12)
M	Y	Y	pleckstrin homology-like domain, family A, member 2	PHLDA2 (Phlda2)
P	?	Y	pleiomorphic adenoma gene-like 1	PLAGL1 (Plagl1)
M	Y	Y	protein phosphatase 1, regulatory (inhibitor) subunit 9A	PPP1R9A (Ppp1r9a)
M	?	Y	primase, DNA, polypeptide 2 (58kDa)	PRIM2 (Prim2)
P	Y	?	Ras protein-specific guanine nucleotide-releasing factor 1	RASGRF1 (Rasgrf1)
P	Y	Y	sarcoglycan, epsilon	SGCE (Sgce)
M	Y	Y	solute carrier family 22, member 18	SLC22A18 (Slc22a18)
M	Y	?	solute carrier family 22 (organic cation transporter), member 2	SLC22A2 (Slc22a2)
M	Y	?	solute carrier family 22 (extraneuronal monoamine transporter), member 3	SLC22A3 (Slc22a3)
P	Y	?	solute carrier family 38, member 4	SLC38A4 (Slc38a4)
P	Y	Y	small nuclear ribonucleoprotein polypeptide N; SNRPN upstream reading frame	SNURF-SNRPN
M	N	Y	transcription elongation factor B polypeptide 3C-like;	TCEB3C
M	Y	Y	tissue factor pathway inhibitor 2	TFPI2 (Tfpi2)
M	?	Y	tumor protein p73	TP73 (Trp73)
P	N	?	transient receptor potential cation channel, subfamily M, member 5	TRPM5 (Trpm5)
M	Y	Y	ubiquitin protein ligase E3A	UBE3A (Ube3a)
P	Y	?	ubiquitin specific peptidase 29	USP29 (Usp29)
P	?	Y	Wilms tumor 1	WT1-Alt transcript (Wt1)
M	Y	no ortholog	zinc finger, imprinted 1	(Zim1)
P/M	Y	Y	paternally expressed 3; PEG3 antisense RNA (non-protein coding); zinc finger, imprinted 2	ZIM2 (Zim2)
M	Y	?	zinc finger, imprinted 3	ZIM3 (Zim3)
P	Y	?	zinc finger protein 264	ZNF264 (Zfp264)
M	?	Y	zinc finger protein 331	ZNF331
M	?	Y	zinc finger protein 597	ZNF597 (Zfp597)

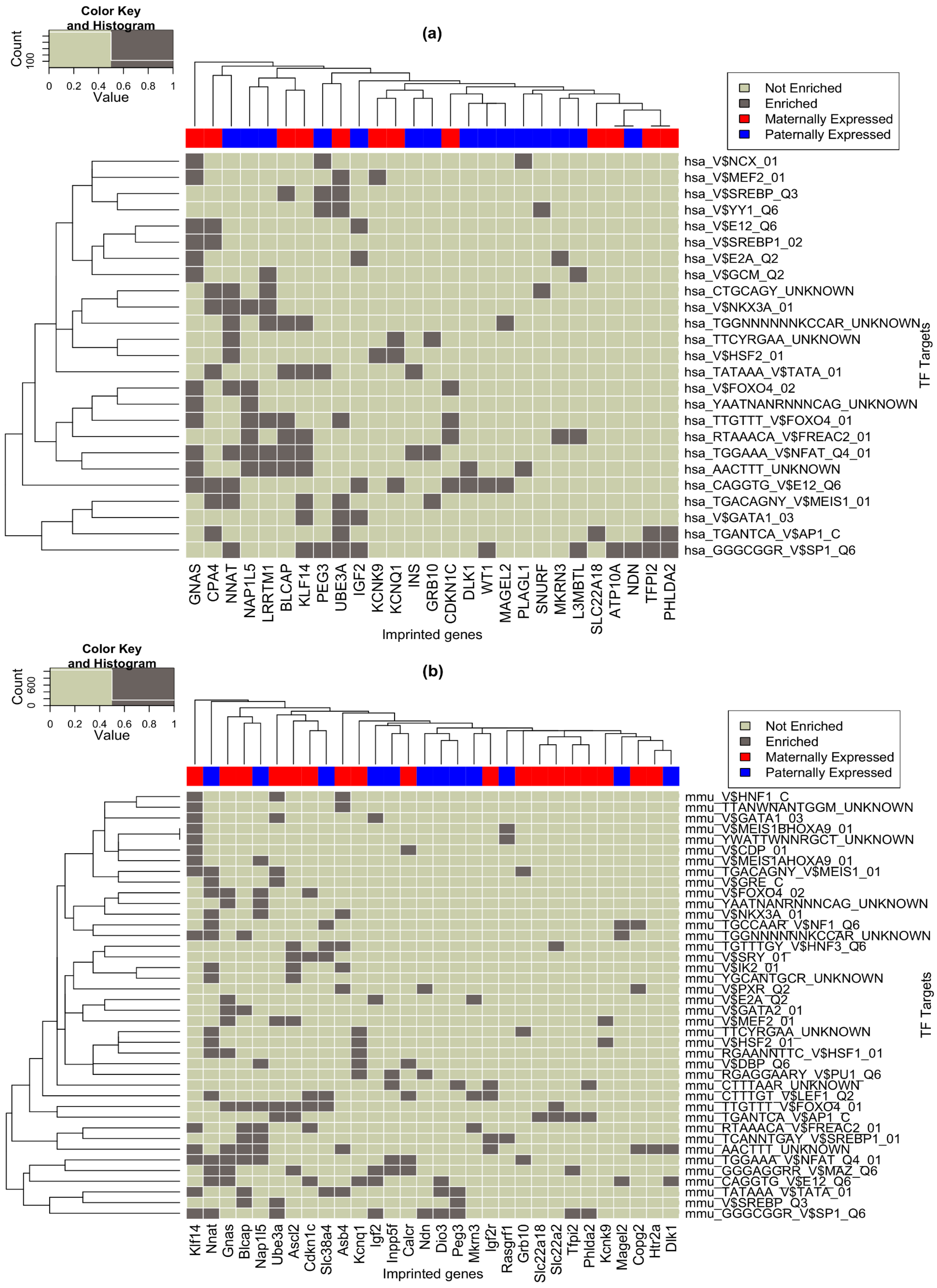


Figure A-1: Heat map for the enriched transcription factor targets in the full set of imprinted genes in human (a) and mouse (b) at p-value of 0.01. Marked in red and blue in the top line are the maternally and paternally expressed genes, respectively.

Appendix B: Supplementary of Chapter 5

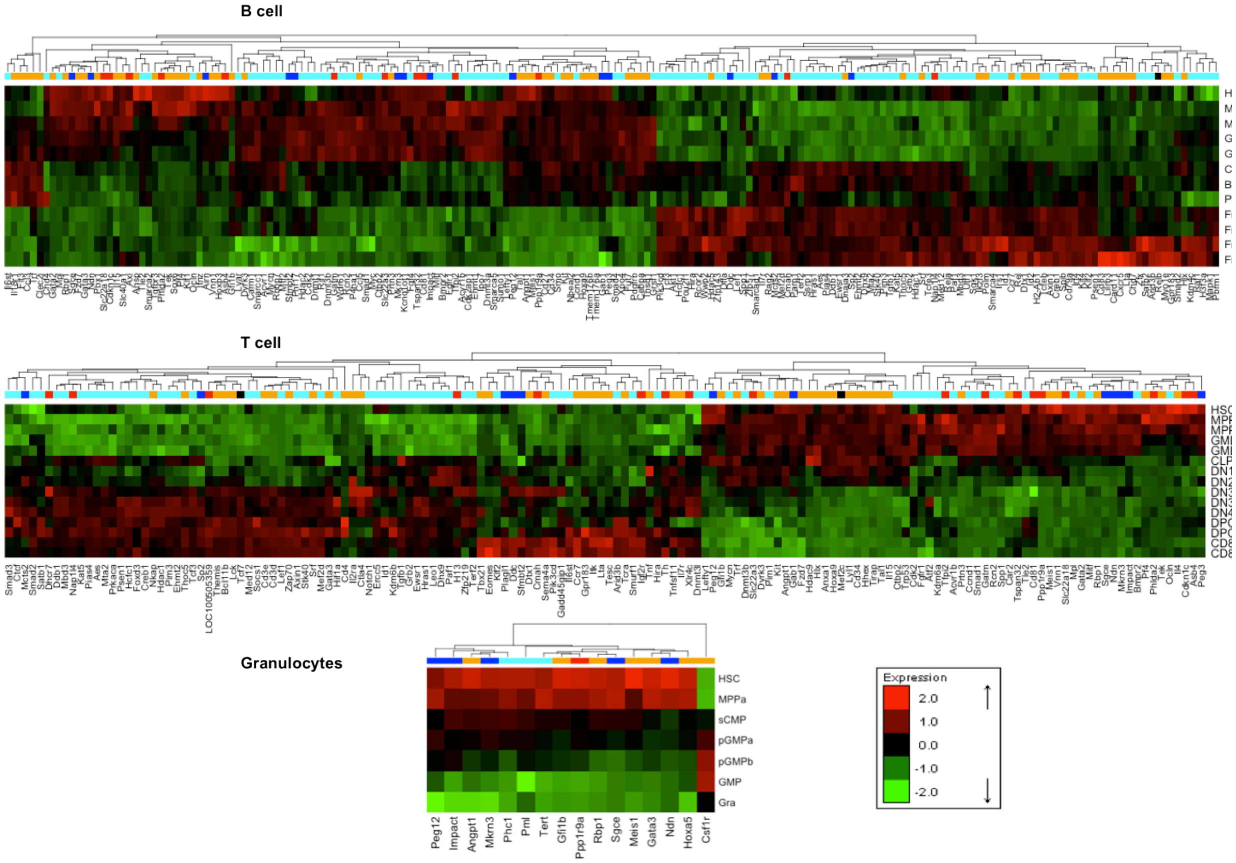


Figure B-1 Heatmaps of differentially expressed imprinted genes (paternally expressed are in blue and maternally expressed are in red), pluripotency genes (cyan), and hematopoietic genes (orange) along three blood lineages (B cells, T cells, and granulocytes) based on GSE34723 dataset. Shared genes between pluripotency and hematopoietic gene sets are marked in black. Green spots represent down-regulated genes, and red spots represent up-regulated genes. The clustering reveals that for every developmental line, there exist imprinted as well pluripotency and hematopoietic genes showing similar expression changes during development.

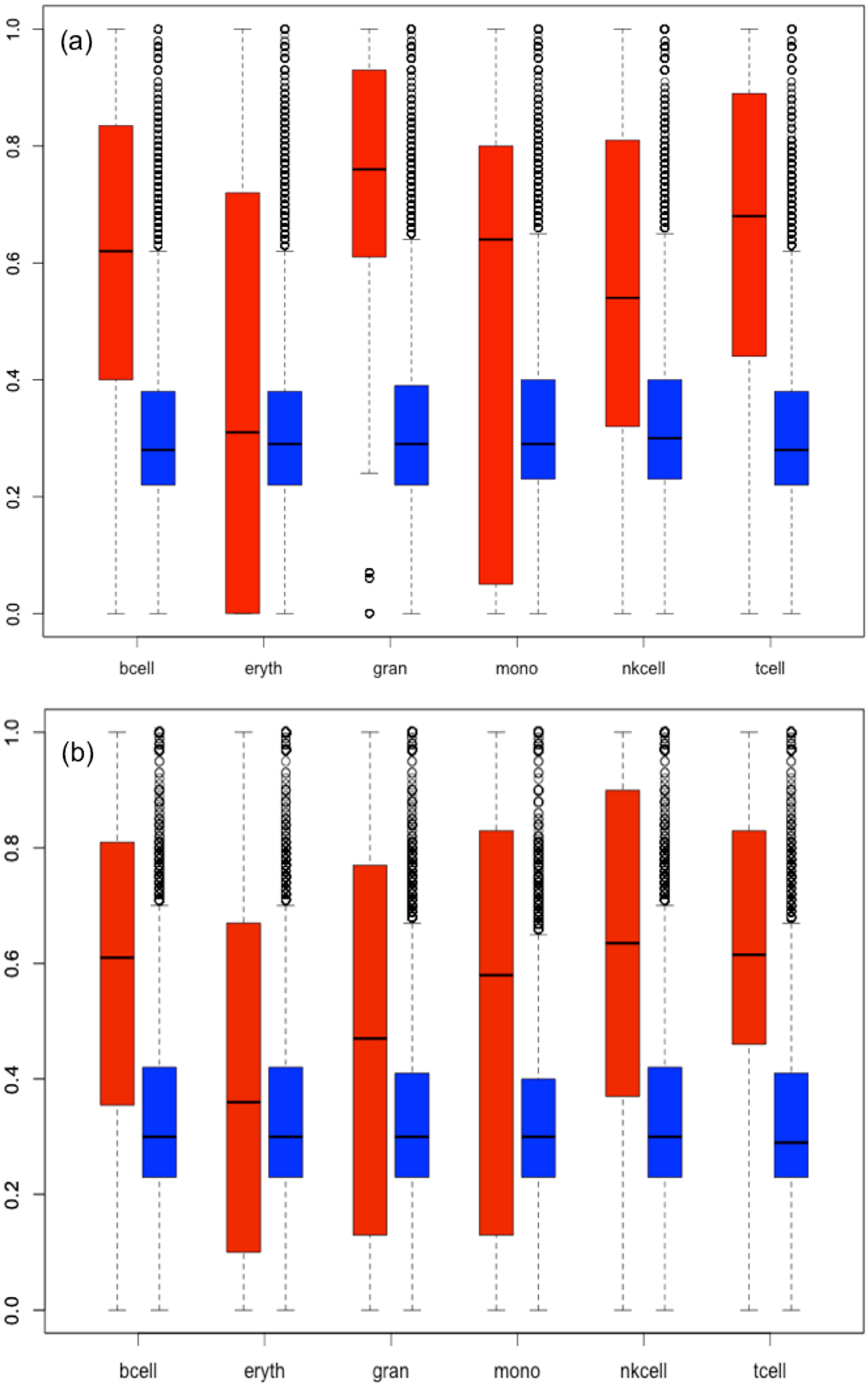


Figure B-2: Functional similarity scores computed with the FunSimMat tool based on the biological process GO category between: A) imprinted and hematopoietic differentially expressed genes (red) compared to the similarity of the other genes in both gene sets (blue) and B) imprinted and differentially expressed pluripotency genes (red) compared to the similarity of the other genes in both gene sets (blue). In A the differentially expressed imprinted and hematopoietic genes show a significantly higher average functional similarity (~0.35 to 0.75) to each other than the background of the other genes in the two gene sets (about 0.3). P-values vary between 0.178 and 6.0 E-237. In B the deferentially expressed imprinted and pluripotency genes show a significantly higher average functional similarity (~0.38 to 0.64) to each other than the background (about 0.3). P-values vary between 0.006 and 4.5 E-24.

Table B-1. List of lineage-specific imprinted, pluripotency, and hematopoietic genes in the investigated blood lineages and the associated mammalian phenotypes due to gene knock outs according to the MGI database.

In order to backup the postulated functional role of the identified lineage markers during hematopoiesis, we checked the MGI database for the mammalian phenotypes associated with abnormalities of hematopoiesis after knocking out these gene alleles. This table lists important hematopoiesis-related phenotypes that are associated with each lineage according to the MGI database. Apparently, multiple lineage-specific genes show deficiencies in either functionalities or differentiation of a lineage, validating the used approach in identifying the lineage markers. An example from the B-cell lineage is the knockout of the imprinted gene CD81. This is reported to cause abnormal B cell morphology (MGI ID: MP:0004939), decreased B-1 B cell number (MP:0004978), and instability in B cell proliferation (MP:0005154, MP:0005093). More generally, the knockout of the imprinted gene Cdkn1c leads to decreasing hematopoietic stem cell number (MP:0004810) and abnormal hematopoietic stem cell physiology (MP:0010763). From the set of pluripotency genes, gene knockout of Relb exhibits also several abnormalities such as decreased B cell number (MP:0005017), decreased B cell proliferation (MP:0005093), absent lymph nodes (MP:0008024), decreased pre-B cell number (MP:0008209), and extra-medullary hematopoiesis (MP:0000240).

Lineage	Imprinted genes count	Lineage-specific imprinted genes	Plurigenes count	Lineage-specific plurigenes	Hematopietic genes count	Lineage-specific hematopietic genes	Lineage related phenotypes due to genes knock-out (Not complete)
Bcell	27	Ppp1r9a, Ndn, Slc22a3, Peg12, Sgce, Gatm, Cdkn1c, Gab1, Cmah, Asb4, Impact, Mkrn3, Tspan32, Phlda2, Cd81, Ddc, Mcts2, Tfpi2, Airn, Kcnq1ot1, Peg3, Sp2, Axl, Sfmbt2, Slc22a18, Nap1l4, Phf17	102	Mpl, Smo, Ccnd1, Bmpr2, Relb, Gab1, Arid3b, Ctbp2, Rel, Tle2, Spp1, Tcf3, Mitf, Tcfeb, Lefty1, Klf2, Akt1, Creb1, Hcfc1, Mef2d, Smad1, Klf4, Ewsr1, Pik3cd, Tgfb1, Irs1, Pou2f1, Lef1, Psen1, Axin1, Rcn2, Dnmt3b, Pim3, Smarca4, Dhx9, Ehmt2, Mta2, Hras1, Kat5, Rif1, Stk40, Raf1, Sgk1, Myc, Zfx, Mbd3, Mapk1, Fgfr1, Hira, Smarca2, Zfp143, Carm1, Parp1, Acvr1b, Xpo4, Smarcad1, Ssrp1, P4ha1, Pias4, Satb2, Id1, Dffa, Paf1, Mycn, Ocln, Pbrm1, Rcor2, Wdr61, Fgf4, Wwp2, H3f3a, Smarcc1, Rbbp7, Grb2, Med12, Mtf2, Dnmt3a, Sumo1, Tcfe3, Ehmt1, Aes, Lyar, Smad4, Cdk2ap1, Il6st, Terf2, Chd4, Kdm4c, Ddb1, Smarca5, Phf17, Zfp57, Hdac1, Rela, Cdk2, Utf1, Hdac2, Grsf1, Ipo7, Smad2, Dnmt1, Acvr1	64	Ccr7, Irf4, Meis1, Cd79a, Tmem176b, Fzd7, Tmem176a, Sox6, Hoxb3, Vnn1, Rbp1, Hoxa9, Ikzf3, Tgfbr3, Nbeal2, Prtn3, Dtx1, Pbx1, Dnaja3, Id2, Cd27, Polm, Pdgfrb, Dyrk3, Ccl5, Il7r, Fut7, Relb, Card11, Thsd1, Myo1e, Klf1, Il15, Rag2, Cxcr5, Slc40a1, Cebpa, Ahsp, Gfi1b, Gpr183, Flt3, Ccl3, Lta, Cd83, Lilrb3, Chd7, Il18r1, Angpt1, Tal1, Gata3, Kit, Spib, Ifnz, Tek, Gata2, H2-Ab1, Hdac5, Cd34, Pf4, Thoc5, Srf, Clec2i, Hlx, Trf	MP:0002144-abnormal B cell differentiation MP:0004939-abnormal B cell morphology MP:0004978-decreased B-1 B cell number MP:0005093-decreased B cell proliferation MP:0008024-absent lymph nodes MP:0008209-decreased pre-B cell number MP:0000702-enlarged lymph nodes MP:0002023-B cell derived lymphoma MP:0002401-abnormal lymphopoiesis MP:0010763-abnormal hematopoietic stem cell physiology MP:0008102-lymph node hyperplasia MP:0004810-decreased hematopoietic stem cell number MP:0010763-abnormal hematopoietic stem cell physiology MP:0002459-abnormal B cell physiology MP:0008174-decreased follicular B cell number MP:0008470-abnormal spleen B cell follicle morphology MP:0005154- increased B cell proliferation MP:0004978- decreased B-1 B cell number
Erythrocytes	4	Sgce, Mkrn3, Kcnq1ot1, Sfmbt2	8	Stat3, Rcn2, Mpl, Satb1, Mef2c, Acvr1b, Smad1, Hras1	11	Tek, Add2, Fli1, Crip2, Rbp1, Gata2, Satb1, Cd27, Ahsp, Mef2c, Acvr1b	MP:0008973-decreased erythroid progenitor cell number MP:0009395-increased nucleated erythrocyte cell number MP:0002875-decreased erythrocyte cell number MP:0000245-abnormal erythropoiesis MP:0002447-abnormal erythrocyte morphology MP:0003656-abnormal erythrocyte physiology MP:0003657-abnormal erythrocyte osmotic lysis MP:0002416-abnormal proerythroblast morphology MP:0003131-increased erythrocyte cell number MP:0003135-increased erythroid progenitor cell number
Granulocytes	6	Ppp1r9a, Sgce, Ndn, Peg12, Impact, Mkrn3	3	Pml, Tert, Phc1	7	Hoxa5, Gfi1b, Gata3, Rbp1, Angpt1, Meis1, Csf1r	MP:0000334-decreased granulocyte number MP:0005072-abnormal hair follicle melanin granule morphology MP:0000322-increased granulocyte number MP:0002396-abnormal hematopoietic system morphology/development MP:0002123-abnormal hematopoiesis MP:0000715-decreased thymocyte number
Monocytes	9	Sgce, Peg12, Ndn, Ppp1r9a, Impact, Klrb1f, Mkrn3, Phlda2, Cdkn1c	6	Mpl, Tle2, Tert, Phc1, Rel, Smad7	19	Cebpa, Csf1r, Egr1, Lgals1, Hoxa5, Gfi1b, Car2, Gata3, Rbp1, Angpt1, Gimap5, Tgfbr3, Pglyrp1, Meis1, Gata2, Sema4a, Nrarp, Tek, Junb	MP:0008112-abnormal monocyte differentiation MP:0002445-abnormal mononuclear cell differentiation MP:0000220-increased monocyte cell number MP:0000223-decreased monocyte cell number MP:0002123-abnormal hematopoiesis
Nkcell	12	Klrb1f, Ppp1r9a, Cdkn1c, Gab1, Ndn, Slc22a3, Sgce, Phlda2, Impact, Mkrn3, Cd81, Ampd3	16	Mpl, Tcf7, Gab1, Relb, Lef1, Smo, Rif1, Mitf, Gatad2a, Klf2, Chd4, Rbl2, Sp1, Atrx, Axin1, Mycn	45	Lck, Rbp1, Fzd7, Ccr2, Cd28, Id2, Card11, Tcf7, Meis1, Txk, Ifng, Vnn1, Tbx21, Sox6, Eomes, Tesc, Cd3d, Ikzf3, Bcl11a, Prdm1, Relb, Lef1,	MP:0008040-decreased NK T cell number MP:0002339-abnormal lymph node morphology MP:0002123-abnormal hematopoiesis MP:0008047-absent uterine NK cells MP:0008038-abnormal NK T cell number

						Angpt1, Tiparp, Kit, Ccl3, Tal1, Gata2, Sema4a, Zap70, Dyrk3, Ccl5, Gab3, Lyl1, Hoxa9, Prtn3, Tgfbr3, Tek, Chd7, Dtx1, H2-Oa, Hdac9, Hlx, Polm, Rsad2	MP:0008044-increased NK cell number MP:0008045-decreased NK cell number MP:0008046-absent NK cells
Tcell	30	Ndn, Ppp1r9a, Sgce, Peg12, Gab1, Asb4, Slc22a3, Mkrn3, LOC100505359, Cdkn1c, Phlda2, Cmah, Gatm, Impact, Igf2r, Tfpi2, Slc22a18, Nap1l4, Sfmbt2, Th, Peg3, Mcts2, Sp2, Dhcr7, Plagl1, Ddc, H13, Tspan32, Cd81, Xlr4c	70	Mpl, Tcf7, Ctbp2, Spp1, Gab1, Mef2c, Kdm6b, Zfp219, Tle2, Smad3, Lefty1, Ccnd1, Rcn2, Satb1, Smad1, Mitf, Lef1, Creb1, Pias4, Psen1, Mef2d, Fgfr1, Stk40, Klf2, Ocln, Socs1, Hras1, Ewsr1, Hdac1, Bmpr2, Mycn, Axin1, Ctcf, Aes, Grb2, Mbd3, Pim1, Ercc5, Hcfc1, Dnmt3b, Ehmt2, Pik3cd, Paf1, Mta2, Dhx9, Terf2, Ddb1, Med12, Gadd45gip1, Pim3, Smurf1, Tgfb1, Arid3b, Cdk2, Hira, Id1, Prkaca, Foxd3, Notch1, Hif1a, Il6st, Leo1, Tcf3, Smad2, Kat5, Acvr1b, Trp53, Atf2, Kdm6a, Dnmt3l	53	Lck, Fzd7, Rbp1, Tcf7, Gata2, Dtx1, Prtn3, Tek, Tnfsf11, Il15, Meis1, Zap70, Bcl11b, Cd3e, Kit, Srf, Cd3d, Vnn1, Car2, Tal1, Dyrk3, Tirap, Lyl1, Pf4, Tesc, Sema4a, Anxa1, Hoxa9, Mef2c, Angpt1, Il7r, Gfi1b, Themis, Lta, Hdac9, Gata3, Itk, Ctla4, Tnf, Cd34, Hhex, Hlx, Gpr183, Ccr7, Cd4, Tcra, Nkap, Thoc5, Il2ra, Trf, Il4, Tbx21, Eomes	MP:0002145-abnormal T cell differentiation MP:0005018-decreased T cell number MP:0008075-decreased CD4-positive T cell number MP:0008079-decreased CD8-positive T cell number MP:0008083-decreased single-positive T cell number MP:0002123-abnormal hematopoiesis MP:0008051-abnormal memory T cell physiology MP:0002024-T cell derived lymphoma MP:0008070-absent T cells
Shared by all Lymphoid lineages: Ppp1r9a, Ndn, Slc22a3, Sgce, Cdkn1c, Gab1, Impact, Mkrn3, Phlda2, Cd81, Meis1, Fzd7, Vnn1, Rbp1, Hoxa9, Prtn3, Dtx1, Dyrk3, Angpt1, Tal1, Kit, Tek, Gata2, Hlx, Mpl, Mitf, Klf2, Lef1, Axin1, Mycn							
Shared by all Myeloid lineages: Sgce, Mkrn3, Rbp1							
Exclusive in Lymphoid Lineages: Slc22a3, Gatm, Gab1, Cmah, Asb4, Tspan32, Cd81, Ddc, Mcts2, Tfpi2, Airn, Peg3, Sp2, Axl, Slc22a18, Nap1l4, Phf17, Ampd3, LOC100505359, Igf2r, Th, Dhcr7, Plagl1, H13, Xlr4c, Ccr7, Irf4, Cd79a, Tmem176b, Fzd7, Tmem176a, Sox6, Hoxb3, Vnn1, Hoxa9, Ikzf3, Nbeal2, Prtn3, Dtx1, Pbx1, Dnaja3, Id2, Polm, Pdgfrb, Dyrk3, Ccl5, Il7r, Fut7, Relb, Card11, Thsd1, Myo1e, Klf1, Il15, Rag2, Cxcr5, Slc40a1, Gpr183, Flt3, Ccl3, Lta, Cd83, Lilrb3, Chd7, Il18r1, Tal1, Kit, Spib, Ifnz, H2-Ab1, Hdac5, Cd34, Pf4, Thoc5, Srf, Clec2i, Hlx, Trf, Lck, Ccr2, Cd28, Tcf7, Txk, Ifng, Tbx21, Eomes, Tesc, Cd3d, Bcl11a, Prdm1, Lef1, Tiparp, Zap70, Gab3, Lyl1, H2-Oa, Hdac9, Rsad2, Tnfsf11, Bcl11b, Cd3e, Tirap, Anxa1, Themis, Itk, Ctla4, Tnf, Hhex, Cd4, Tcra, Nkap, Il2ra, Il4, Smo, Ccnd1, Bmpr2, Arid3b, Ctbp2, Spp1, Tcf3, Mitf, Tcfeb, Lefty1, Klf2, Akt1, Creb1, Hcfc1, Mef2d, Klf4, Ewsr1, Pik3cd, Tgfb1, Irs1, Pou2f1, Psen1, Axin1, Dnmt3b, Pim3, Smarca4, Dhx9, Ehmt2, Mta2, Kat5, Rif1, Stk40, Raf1, Sgk1, Myc, Zfx, Mbd3, Mapk1, Fgfr1, Hira, Smarca2, Zfp143, Carm1, Parp1, Xpo4, Smarcad1, Ssrp1, P4ha1, Pias4, Satb2, Id1, Dffa, Paf1, Mycn, Ocln, Pbrm1, Rcor2, Wdr61, Fgf4, Wwp2, H3f3a, Smarcc1, Rbbp7, Grb2, Med12, Mtf2, Dnmt3a, Sumo1, Tcfe3, Ehmt1, Aes, Lyar, Smad4, Cdk2ap1, Il6st, Terf2, Chd4, Kdm4c, Ddb1, Smarca5, Zfp57, Hdac1, Rela, Cdk2, Utf1, Hdac2, Grsf1, Ipo7, Smad2, Dnmt1, Acvr1, Gatad2a, Rbl2, Sp1, Atrx, Kdm6b, Zfp219, Smad3, Socs1, Ctcf, Pim1, Ercc5, Gadd45gip1, Smurf1, Prkaca, Foxd3, Notch1, Hif1a, Leo1, Trp53, Atf2, Kdm6a, Dnmt3l							
Exclusive in Myeloid Lineages: Add2, Fli1, Crip2, Hoxa5, Csf1r, Egr1, Lgals1, Gimap5, Pglyrp1, Nrarp, Junb, Stat3, Pml, Tert, Phc1, Smad7							

Table B-2. Gene sets studied in this work.

Gene set	Count	Annotated in MS4302.0 array	Description
Imprinted genes	*120*	*86*	*Imprinted genes selected from the Imprinting catalogs as described in methods*
Pluripotency genes	*274*	*272*	*Genes involved in the PluriNetwork*
Hematopoietic genes	*615*	*562*	*Genes annotated for GO:0048534: "hematopoietic or lymphoid organ development"*
Ign genes	*169*	*155*	*Imprinted genes plus additional genes regulating the imprinted genes*
Ignpluri genes	*20*	*20*	*Genes involved in both Ign and PluriNetwork*
Ignhema genes	*17*	*17*	*Genes involved in both Ign and hematopoietic genes*
Ignshared genes	*32*	*32*	*Combined list of genes shared between 1- Ign and hematopoietic genes. And 2- Ign and pluripotency genes*
Gene population (Total number of genes in the array) = 21390			

Appendix C: Supplementary of Chapter 6

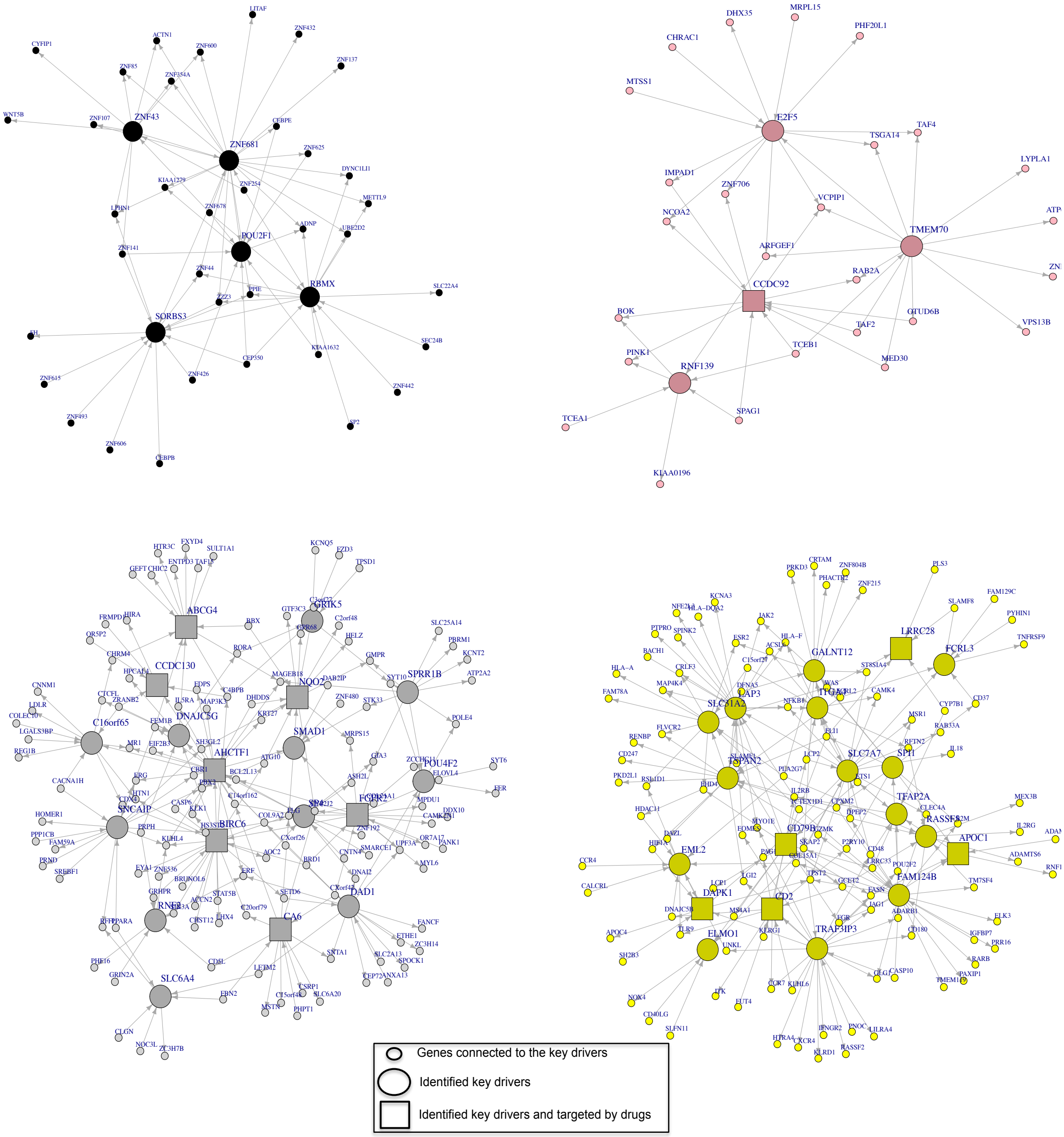


Figure C-1. The inferred regulatory networks for the black, pink, grey, and yellow gene modules. For clarity, we visualized only the identified key driver genes and the nodes connected to them.

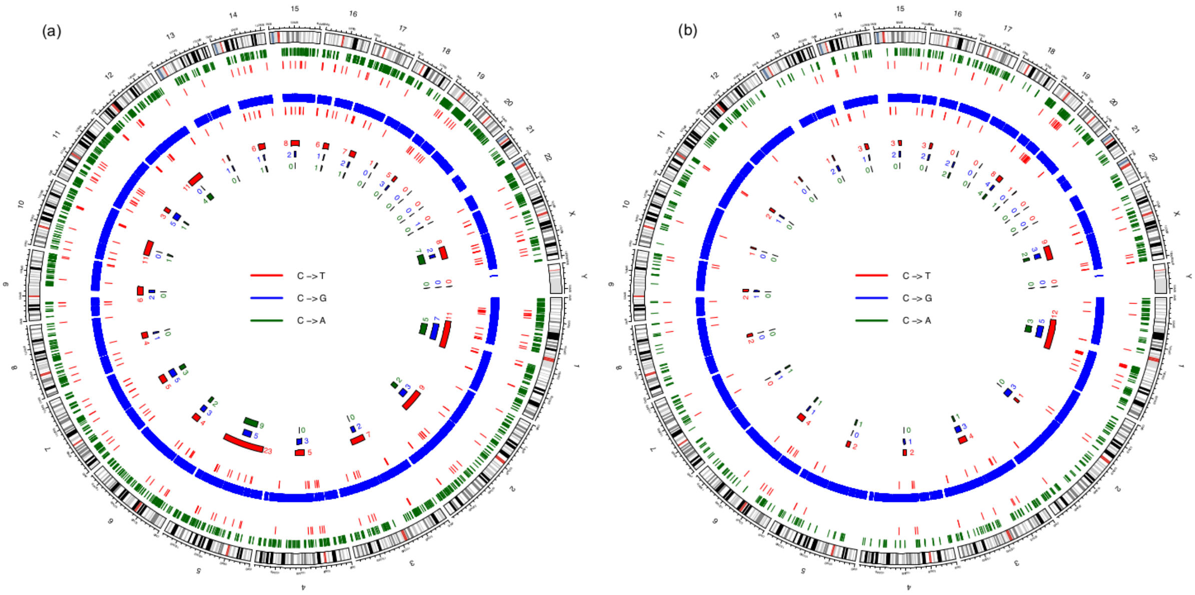


Figure C-2. Proximity analysis of somatic mutations with the up-and down-methylated genes. Ideogram plots showing the genomic distributions of the somatic mutations occurring at promoter regions of (a) the up-methylated genes (234 cases), and (b) down-methylated genes (113 cases). The outer green circle shows the entire set of differentially methylated genes, whereas the next highlighted red lines refer to the identified cases adjacent to the somatic mutations. The inner blue circle represents the entire set of somatic SNVs and the next highlighted red lines depict the matched SNVs in the identified cases. The plot illustrates also the fractions of the three considered types of mutations (C->T, C->G and C->A) showing the occurrence frequency for each. Obviously the C->T mutations for the up-methylated genes occur at a higher rate than its peers in the down-methylated genes.

Table C-1. Ten most significant GO terms and KEGG pathways enriched in the list of the 73 candidate driver genes.

Category	Enriched term	P-value
GO functional terms	GO:0006357~regulation of transcription from RNA polymerase II promoter	6.67E-09
	GO:0006355~regulation of transcription, DNA-dependent	1.15E-07
	GO:0006350~transcription	1.59E-07
	GO:0051252~regulation of RNA metabolic process	1.75E-07
	GO:0045449~regulation of transcription	1.96E-07
	GO:0034645~cellular macromolecule biosynthetic process	1.08E-06
	GO:0019219~regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.10E-06
	GO:0010556~regulation of macromolecule biosynthetic process	1.24E-06
	GO:0009059~macromolecule biosynthetic process	1.26E-06
	GO:0051171~regulation of nitrogen compound metabolic process	1.33E-06
KEGG pathways	hsa05223:Non-small cell lung cancer	2.48E-03
	hsa04110:Cell cycle	3.42E-03
	hsa05215:Prostate cancer	1.01E-02
	hsa05219:Bladder cancer	1.91E-02
	hsa05200:Pathways in cancer	2.32E-02
	hsa05214:Glioma	4.06E-02

Table C-2. A list of the 33 genes whose gene products are targeted by anti-cancer drugs, characterized from the three considered drug databases, CTD, PharmGKB, and Cancer resource. (1) means that at least one drug that targets this gene product is reported in this database, and (0) means no drugs are reported for the respective gene in this database. Not included are substances that are known to be cancerogenous or mutagenic.

Target gene	Drug and antineoplastic agents	CTD	PharmGKB	Cancer Resource
ABCB8	docetaxel; Cyclosporine; Progesterone	1	0	0
ABCG4	indole-3 carbinol; Methotrexate; exemestane; Vincristine	1	0	0
AHCTF1	Methotrexate; bisphenol A	1	0	0
AKT1	U 0126;tyrphostin AG 1478; Ursodeoxycholic Acid;Valproic Acid;tyrphostin AG 1024; trametinib; Tretinoin	1	0	1
APOC1	tanshinone; Quercetin; Fluorouracil; bexarotene; Cisplatin; Tamoxifen	1	0	1
AR	Dihydrotestosterone; Acetylcysteine; celecoxib	1	0	0
ATF6	Nelfinavir; Tretinoin;bisphenol A; Cyclosporine; Curcumin	1	0	0
ATG4C	epigallocatechin gallate	1	0	0
ATP1B1	resveratrol; Ranitidine; vorinostat; Genistein; Progesterone; epigallocatechin gallate	1	0	0
B4GALT7	Cytarabine; Cyclosporine	1	0	0
BIRC6	Dieldrin; Cyclosporine; Cisplatin; Fluorouracil; Doxorubicin; Epirubicin;Estradiol; zoledronic acid; bisphenol A	1	0	0
BRCA1	Tretinoin; trichostatin A; Estradiol; transplatin; troglitazone; Tunicamycin; fulvestrant	1	0	1
CA6	Tretinoin;Carmustine	1	0	0
CCDC130	Quercetin;Tamoxifen;Cyclosporine;bisphenol A	1	0	0
CCDC92	Quercetin; Folic Acid	1	0	0
CD2	Dexamethasone; Methotrexate; Cyclophosphamide	1	0	0
CD79B	Cyclophosphamide	1	0	0
CDC34	Estradiol; bortezomib; Fluorouracil; Tamoxifen	1	0	0
DAPK1	paclitaxel;gemcitabine	0	1	0
EGR1	Fluorouracil; gemcitabine	0	0	1
ESR1	exemestane;tamoxifen	0	1	1
JUN	andrographolide; cinnamic aldehyde; Daunorubicin; decitabine; Cisplatin;Doxorubicin	0	0	1
LRRC28	gemcitabine	0	0	1
MYB	Fluorouracil;gemcitabine;Quercetin	0	0	1
MYC	alitretionoin; Amsarcine; bicalutamide; Camtothecin; decitabine; Cisplatin; Doxorubicin	0	0	1
NFKB1	Curcumin; decitabine; Doorubicin; Echinomycin; Fluorouracil; gefitinib; indole-3-carbinol; parthenolide	0	0	1
NQO2	doxorubicin; cyclophosphamide	0	1	0
OS9	alitretionoin	0	0	1
SP1	Etoposide; indole-3-carbinol; Ionidamine; Quercetin; Adaphostin	0	0	1
STAT3	azaspirane; bisphenol A; Capsaicin; Fluorouracil; interferon alfacon-1; resveratrol;sulindac sulfide; Tamoxifen	0	0	1
TGFB1	Doxorubicin; Fluorouracil; Thalidomide; Entinostat; Hyaluronidase	0	0	1
TP53	4-biphenylmine; alliin; Apigenin; Atropine;bicalutamide;butylidenephthalide	0	0	1

Appendix D: Supplementary of Chapter 7

Table D-1: Final list of 58 GRN module genes and transcription factors.

Symbol	Entrez Gene ID	Definition	LFC (ratio)	P-value	Regulation
ABCC2	12780	Mus musculus ATP-binding cassette, sub-family C (CFTR/MRP), member 2 (Abcc2), mRNA.	0.9920	0.0111	DOWN
AHR	11622	Mus musculus aryl-hydrocarbon receptor (Ahr), mRNA.	1.0029	0.3787	UP
ALX1	216285	Mus musculus ALX homeobox 1 (Alx1), mRNA.	1.0025	0.2549	UP
AP1S1	11769	Mus musculus adaptor protein complex AP-1, sigma 1 (Ap1s1), mRNA.	0.9886	0.3645	DOWN
AR	11835	Mus musculus androgen receptor (Ar), mRNA.	1.0027	0.4901	UP
ASPSCR1	68938	Mus musculus alveolar soft part sarcoma chromosome region, candidate 1 (human) (Aspscr1), transcript variant 2, mRNA.	1.0093	0.0419	UP
BMF	NA	NA	1.0324	0.0001	UP
CASP3	NA	NA	0.9910	0.0208	DOWN
CCT7	12468	Mus musculus chaperonin containing Tcp1, subunit 7 (eta) (Cct7), mRNA.	0.9792	0.0020	DOWN
CDS1	74596	Mus musculus CDP-diacylglycerol synthase 1 (Cds1), mRNA.	0.9952	0.7725	DOWN
CEBPA	12606	Mus musculus CCAAT/enhancer binding protein (C/EBP), alpha (Cebpa), mRNA.	0.9868	0.3237	DOWN
CLCNKA	12733	Mus musculus chloride channel Ka (Clcnka), mRNA.	0.9908	0.0089	DOWN
DR1	13486	Mus musculus down-regulator of transcription 1 (Dr1), mRNA.	0.9984	0.7765	DOWN
FOXA1	15375	Mus musculus forkhead box A1 (Foxa1), mRNA.	0.9948	0.1571	DOWN
FOXA2	15376	Mus musculus forkhead box A2 (Foxa2), mRNA.	0.9991	0.6021	DOWN
FOXD3	15221	Mus musculus forkhead box D3 (Foxd3), mRNA.	1.0019	0.5498	UP
FOXF2	14238	Mus musculus forkhead box F2 (Foxf2), mRNA.	0.9998	0.9395	DOWN
FOXI1	14233	Mus musculus forkhead box I1 (Foxi1), mRNA.	1.0043	0.0695	UP
FOXJ1	15223	Mus musculus forkhead box J1 (Foxj1), mRNA.	0.9964	0.2217	DOWN
FOXJ2	60611	Mus musculus forkhead box J2 (Foxj2), mRNA.	1.0160	0.1175	UP
FOXL1	14241	Mus musculus forkhead box L1 (Foxl1), mRNA.	1.0023	0.4621	UP
FOXO1	NA	NA	1.0125	0.3555	UP
FOXO4	54601	Mus musculus forkhead box O4 (Foxo4), mRNA.	1.0003	0.9086	UP
FOXQ1	15220	Mus musculus forkhead box Q1 (Foxq1), mRNA.	1.0081	0.3548	UP
GATA1	14460	Mus musculus GATA binding protein 1 (Gata1), mRNA.	0.9974	0.8083	DOWN
HNF4A	15378	Mus musculus hepatic nuclear factor 4, alpha (Hnf4a), mRNA.	0.9990	0.7515	DOWN
HNF4G	30942	Mus musculus hepatocyte nuclear factor 4, gamma (Hnf4g), mRNA.	1.0041	0.1770	UP
IKZF1	22778	Mus musculus IKAROS family zinc finger 1 (Ikzf1), transcript variant 1, mRNA.	1.0018	0.8153	UP
MAX	17187	Mus musculus Max protein (Max), mRNA.	1.0166	0.3979	UP
MAZ	17188	Mus musculus MYC-associated zinc finger protein (purine-binding transcription factor) (Maz), mRNA.	0.9978	0.4452	DOWN
MEF2C	NA		1.0114	0.6041	UP
NFATC2	18019	Mus musculus nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2 (Nfatc2), transcript variant 2, mRNA.	0.9993	0.6261	DOWN
NFE2	18022	Mus musculus nuclear factor, erythroid derived 2 (Nfe2), mRNA.	1.0174	0.1184	UP
NFKB1	NA		1.0048	0.6109	UP
NFYA	18044	Mus musculus nuclear transcription factor-Y alpha (Nfya), mRNA.	0.9975	0.5666	DOWN
NKX6-2	14912	Mus musculus NK6 homeobox 2 (Nkx6-2), mRNA.	0.9985	0.4802	DOWN
NR2F2	11819	Mus musculus nuclear receptor subfamily 2, group F, member 2 (Nr2f2), transcript variant 2, mRNA.	0.9998	0.8905	DOWN

PIK3C2A	NA	NA	1.0080	0.0003	UP
POU6F1	19009	Mus musculus POU domain, class 6, transcription factor 1 (Pou6f1), mRNA.	1.0155	0.0961	UP
PPARA	19013	Mus musculus peroxisome proliferator activated receptor alpha (Ppara), mRNA.	1.0015	0.5671	UP
PPARG	19016	Mus musculus peroxisome proliferator activated receptor gamma (Pparg), mRNA.	0.9966	0.2365	DOWN
PTF1A	NA	NA	1.0064	0.0099	UP
RUNX1	12394	Mus musculus runt related transcription factor 1 (Runx1), mRNA.	0.9954	0.5370	DOWN
SLC12A1	NA	NA	1.0026	0.0054	UP
SLC16A4	229699	Mus musculus solute carrier family 16 (monocarboxylic acid transporters), member 4 (Slc16a4), mRNA.	0.9898	0.0008	DOWN
SLC22A1	20517	Mus musculus solute carrier family 22 (organic cation transporter), member 1 (Slc22a1), mRNA.	1.0006	0.7509	UP
SPI15	NA	NA	1.0020	0.4474	UP
SREBF1	20787	Mus musculus sterol regulatory element binding factor 1 (Srebf1), mRNA.	0.9965	0.2896	DOWN
SRY	21674	Mus musculus sex determining region of Chr Y (Sry), mRNA.	1.0003	0.9368	UP
TBP	21374	Mus musculus TATA box binding protein (Tbp), mRNA.	0.9860	0.0673	DOWN
TCF3	21415	Mus musculus transcription factor 3 (Tcf3), transcript variant 1, mRNA.	1.0010	0.6035	UP
TCF7	21414	Mus musculus transcription factor 7, T-cell specific (Tcf7), mRNA.	0.9992	0.9032	DOWN
TEF	21685	Mus musculus thyrotroph embryonic factor (Tef), transcript variant 1, mRNA.	1.0029	0.4737	UP
VDR	22337	Mus musculus vitamin D receptor (Vdr), mRNA.	1.0022	0.6336	UP
ZEB1	21417	Mus musculus zinc finger E-box binding homeobox 1 (Zeb1), mRNA.	1.0110	0.4390	UP
E4BP4	NA	Unannotated in the microarray chip	NA	NA	NA
TCF11MAFG	NA	Unannotated in the microarray chip	NA	NA	NA
AHRARNT	NA	Unannotated in the microarray chip	NA	NA	NA

Appendix E: Supplementary of Chapter 8

Table E-1. List of the twice-mutated genes and the associated functional terms.

Gene Name	Gene Symbol	Description	Functional / metabolic /protein family group
SA2981_1390	**ebhB**	Putative Staphylococcal surface anchored protein; adhesin emb	N.A
SA2981_1815	**pfoS/R**	Regulatory protein	N.A
SA2981_1256	**glpF**	Glycerol uptake facilitator protein	N.A
SA2981_2486	**feoB**	Ferrous iron transport protein B	N.A
SA2981_2564	**yvcP**	Two-component response regulator YvcP	N.A
SA2981_0120	**sbnD**	Siderophore staphylobactin biosynthesis protein SbnD	Tetracycline resistance protein, TetA (INTERPRO) tetracycline transport (GO:0015904) antibiotic transport (*GO:0042891*) drug:hydrogen antiporter activity (GO:0015307) response to stimulus (GO:0050896)
SA2981_1100	**mutS2**	Recombination inhibitory protein MutS2	response to stimulus (GO:0050896)
SA2981_1178	**prkC**	Serine/threonine protein kinase PrkC, regulator of stationary phase	N.A
SA2981_1260	**miaA**	tRNA delta(2)-isopentenylpyrophosphate transferase	N.A
SA2981_1284	**thrC**	Threonine synthase	N.A
SA2981_1323	**trpD**	Anthranilate phosphoribosyltransferase	Two-component system (KEGG)
SA2981_1468	**gnd**	6-phosphogluconate dehydrogenase, decarboxylating	N.A
SA2981_1511	**sodA**	Manganese superoxide dismutase; Superoxide dismutase (Fe)	response to stimulus (GO:0050896)
SA2981_1826	**tagG**	Teichoic acid translocation permease protein TagG	N.A
SA2981_2019	**kdpD**	Osmosensitive K+ channel histidine kinase KdpD	Two-component system (KEGG:)
SA2981_2265	**metT**	Methionine transporter MetT	N.A
SA2981_2294	**tcaB**	Teicoplanin resistance associated membrane protein TcaB	Tetracycline resistance protein, TetA (INTERPRO:) tetracycline transport (GO:0015904) antibiotic transport (*GO:0042891*) drug:hydrogen antiporter activity (GO:0015307) response to stimulus (GO:0050896)
SA2981_2400	**opp-1F**	Oligopeptide transporter putative ATPase domain protein	N.A

Table E-2. List of the SNPs occurring in at least 2 invasive strains but in none of the nasal strains, their genes, and the resulting amino acid change. For the amino acid change in the five cases of insertions, the original reading frame (ORF) is shifted leading to a wide change in the amino acid chain.

Locus tag	Gene name	Description	SNP position	Refe-rence NT	Alter-native NT	Amino acid change	P-value of the variant
SA2981_0120	sbnD	Siderophore staphylobactin biosynthesis protein SbnD	131858	G	T	W to C	4.05E-51
SA2981_0148	-		162408	T	C	none (D)	2.21E-50

SA2981_0542	-		616048	A	G	none (G)	2.83E-52
SA2981_0561	-		631881	G	A	R to K	3.54E-53
SA2981_0710	-		783925	T	C	none (I)	1.36E-48
SA2981_0711	-		784904	A	G	Q to R	2.83E-61
SA2981_0724	-		797606	G	GA	ORF shifted	1.81E-47
SA2981_0874	-		921757	G	A	G to D	2.97E-53
SA2981_0978	-		1035879	A	G	I to V	4.15E-40
			1036742	G	A	M to I	1.86E-70
SA2981_1074	-		1131989	C	T	T to I	3.51E-44
SA2981_1100	mutS2	Recombination inhibitory protein MutS2	1157221	C	T	none (S)	9.61E-67
SA2981_1178	prkC	Serine/threonine protein kinase PrkC, regulator of stationary phase	1237441	G	T	A to S	2.81E-52
SA2981_1251	-		1324647	C	A	T to N	2.45E-37
SA2981_1256	glpF	Glycerol uptake facilitator protein	1331327	C	CT	ORF shifted	2.41E-29
SA2981_1260	miaA	tRNA delta(2)-isopentenylpyrophosphate transferase	1336837	A	G	D to G	8.05E-45
SA2981_1284	thrC	Threonine synthase	1361487	C	T	S to F	1.61E-30
SA2981_1288	-		1365826	T	C	T to A	1.59E-26
SA2981_1323	trpD	Anthranilate phosphoribosyltransferase	1409763	T	C	R to R	2.78E-38
SA2981_1390	ebhB	Putative Staphylococcal surface anchored protein; adhesin emb	1503065	C	T	S to N	7.67E-41
SA2981_1468	gnd	6-phosphogluconate dehydrogenase, decarboxylating	1585512	G	A	none (F)	2.40E-23
SA2981_1511	sodA	Manganese superoxide dismutase; Superoxide dismutase (Fe)	1622766	C	T	G to D	2.46E-33
SA2981_1815	pfoS/R	Regulatory protein	1946895	C	CAAT	add R	1.93E-31
SA2981_1826	tagG	Teichoic acid translocation permease protein TagG	1967972	A	T	F to Y	2.88E-51
SA2981_2019	kdpD	Osmosensitive K+ channel histidine kinase KdpD	2151346	G	GA	ORF shifted	1.37E-53
SA2981_2265	metT	Methionine transporter MetT	2394447	A	T	none (G)	5.88E-62
SA2981_2284	-		2414503	G	A	A to T	7.91E-77
SA2981_2294	tcaB	Teicoplanin resistance associated membrane protein TcaB	2424068	A	G	I to T	1.70E-62
SA2981_2329	-		2462332	C	T	none (L)	1.14E-59
SA2981_2366	-		2504026	G	A	P to S	8.70E-67
SA2981_2367	-		2504949	A	G	I to V	3.05E-58
SA2981_2370	-		2507139	C	CA	ORF shifted	1.69E-24
SA2981_2400	opp-1F	Oligopeptide transporter putative ATPase domain protein	2541624	C	T	A to T	4.11E-76
SA2981_2486	feoB	Ferrous iron transport protein B	2637050	C	T	V to M	3.31E-61
SA2981_2556	-		2710563	T	C	none (K)	2.77E-50
SA2981_2564	yvcP	Two-component response regulator YvcP	2721091	G	T	P to T	2.38E-71
SA2981_2642	-		2812721	G	T	R to L	3.21E-48

Bibliography

1. Filkov, V., *Identifying Gene Regulatory Networks from Gene Expression Data, chapter 27.* Handbook of Computational Molecular Biology, 2001.
2. Levine, A.J. and M. Oren, *The first 30 years of p53: growing ever more complex.* Nature Reviews Cancer, 2009. **9**(10): p. 749-758.
3. Cuccato, G., G. Della Gatta, and D. di Bernardo, *Systems and Synthetic biology: tackling genetic networks and complex diseases.* Heredity, 2009. **102**(6): p. 527-532.
4. Vijesh, N., S.K. Chakrabarti, and J. Sreekumar, *Modeling of gene regulatory networks: A review.* Journal of Biomedical Science and Engineering, 2013. **6**(02): p. 223.
5. McAdams, H.H. and A. Arkin, *Simulation of prokaryotic genetic circuits.* Annual review of biophysics and biomolecular structure, 1998. **27**(1): p. 199-224.
6. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000. **407**(6804): p. 651-654.
7. Mohamed, H., et al., *Integrative network-based approach identifies key genetic elements in breast invasive carcinoma.* BMC Genomics, 2015. **16**(Suppl 5): p. S2.
8. Davidson, E. and M. Levin, *Gene regulatory networks.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(14): p. 4935-4935.
9. Bird, A., *Perceptions of epigenetics.* Nature, 2007. **447**(7143): p. 396-398.
10. Varriale, A., *DNA methylation, epigenetics, and evolution in vertebrates: facts and challenges.* International journal of evolutionary biology, 2014. **2014**.
11. Simmons, D., *Epigenetic influence and disease.* Nature Education, 2008. **1**(1): p. 6.
12. Jones, P.A., et al., *Moving AHEAD with an international human epigenome project.* Nature, 2008. **454**(7205): p. 711-715.
13. Zhong, C., et al., *Mutations and CpG islands among hepatitis B virus genotypes in Europe.* BMC Bioinformatics, 2015. **16**(1): p. 38.
14. Gu, Y., et al., *Global DNA methylation and transcriptional analyses of human ESC-derived cardiomyocytes.* Protein & cell, 2014. **5**(1): p. 59-68.
15. Das, P.M. and R. Singal, *DNA methylation and cancer.* Journal of Clinical Oncology, 2004. **22**(22): p. 4632-4642.
16. Esteller, M. and J.G. Herman, *Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours.* The Journal of pathology, 2002. **196**(1): p. 1-7.
17. Portela, A. and M. Esteller, *Epigenetic modifications and human disease.* Nature biotechnology, 2010. **28**(10): p. 1057-1068.
18. Li, S., et al., *An optimized algorithm for detecting and annotating regional differential methylation.* BMC bioinformatics, 2013. **14**(Suppl 5): p. S10.
19. Esteller, M., *Cancer epigenomics: DNA methylomes and histone-modification maps.* Nature Reviews Genetics, 2007. **8**(4): p. 286-298.
20. Tahara, T., et al., *Effect of promoter methylation of multidrug resistance 1 (MDR1) gene in gastric carcinogenesis.* Anticancer research, 2009. **29**(1): p. 337-341.
21. Xia, J., L. Han, and Z. Zhao, *Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome.* BMC genomics, 2012. **13**(Suppl 8): p. S7.

22. Herrick, G. and J. Seger, *Imprinting and paternal genome elimination in insects*, in *Genomic Imprinting*. 1999, Springer. p. 41-71.
23. Liu, J., et al., *A GNAS1 imprinting defect in pseudohypoparathyroidism type IB.* Journal of Clinical Investigation, 2000. **106**(9): p. 1167-1174.
24. Lefebvre, L., et al., *Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene Mest.* Nature genetics, 1998. **20**: p. 163-170.
25. Rodenhiser, D. and M. Mann, *Epigenetics and human disease: translating basic biology into clinical applications.* Canadian Medical Association Journal, 2006. **174**(3): p. 341-348.
26. Day, D. and M.F. Tuite, *Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview.* Journal of Endocrinology, 1998. **157**(3): p. 361-371.
27. Alberts, B., *Molecular biology of the cell*. 4th ed. 2002, New York: Garland Science. xxxiv, 1548 p.
28. Cheadle, C., et al., *Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability.* BMC genomics, 2005. **6**(1): p. 75.
29. Kung, J.T., D. Colognori, and J.T. Lee, *Long noncoding RNAs: past, present, and future.* Genetics, 2013. **193**(3): p. 651-669.
30. Wang, J., et al., *TransmiR: a transcription factor–microRNA regulation database.* Nucleic acids research, 2010. **38**(suppl 1): p. D119-D122.
31. Rodriguez, A., et al., *Identification of mammalian microRNA host genes and transcription units.* Genome research, 2004. **14**(10a): p. 1902-1910.
32. Hu, W. and J. Coller, *What comes first: translational repression or mRNA degradation? The deepening mystery of microRNA function.* Cell research, 2012. **22**(9): p. 1322-1324.
33. Volinia, S., et al., *Reprogramming of miRNA networks in cancer and leukemia.* Genome research, 2010. **20**(5): p. 589-599.
34. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome research, 2009. **19**(1): p. 92-105.
35. Taft, R.J., et al., *Non ‐ coding RNAs: regulators of disease.* The Journal of pathology, 2010. **220**(2): p. 126-139.
36. Esquela-Kerscher, A. and F.J. Slack, *Oncomirs—microRNAs with a role in cancer.* Nature Reviews Cancer, 2006. **6**(4): p. 259-269.
37. Medina, P.P. and F.J. Slack, *microRNAs and cancer: an overview.* Cell cycle, 2008. **7**(16): p. 2485-2492.
38. Yanaihara, N., et al., *Unique microRNA molecular profiles in lung cancer diagnosis and prognosis.* Cancer cell, 2006. **9**(3): p. 189-198.
39. Ryan, B.M., A.I. Robles, and C.C. Harris, *Genetic variation in microRNA networks: the implications for cancer research.* Nature Reviews Cancer, 2010. **10**(6): p. 389-402.
40. Sethupathy, P. and F.S. Collins, *MicroRNA target site polymorphisms and human disease.* Trends in genetics, 2008. **24**(10): p. 489-497.
41. Bhattacharya, A., J.D. Ziebarth, and Y. Cui, *Systematic analysis of microRNA targeting impacted by small insertions and deletions in human genome.* PloS one, 2012. **7**(9): p. e46176.
42. Mendell, J.T. and E.N. Olson, *MicroRNAs in stress signaling and human disease.* Cell, 2012. **148**(6): p. 1172-1187.

43. Bhattacharya, A., J.D. Ziebarth, and Y. Cui, *SomamiR: a database for somatic mutations impacting microRNA function in cancer.* Nucleic acids research, 2012: p. gks1138.
44. Abelson, J.F., et al., *Sequence variants in SLITRK1 are associated with Tourette's syndrome.* Science, 2005. **310**(5746): p. 317-320.
45. Chen, C.-Z., et al., *MicroRNAs modulate hematopoietic lineage differentiation.* science, 2004. **303**(5654): p. 83-86.
46. Li, J.-H., et al., *starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data.* Nucleic acids research, 2013: p. gkt1248.
47. Nifoussi, S.K., *Posttranslational regulation of protein function and stability at the mitochondria and beyond.* 2010.
48. Wang, Y.-C., S.E. Peterson, and J.F. Loring, *Protein post-translational modifications and regulation of pluripotency in human stem cells.* Cell research, 2014. **24**(2): p. 143-160.
49. Mei, Y., et al., *Combinatorial development of biomaterials for clonal growth of human pluripotent stem cells.* Nature materials, 2010. **9**(9): p. 768-778.
50. Barabási, A.-L., N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease.* Nature Reviews Genetics, 2011. **12**(1): p. 56-68.
51. Kauffman, S., *Homeostasis and differentiation in random genetic control networks.* Nature, 1969. **224**: p. 177-178.
52. Akutsu, T., S. Miyano, and S. Kuhara, *Inferring qualitative relations in genetic networks and metabolic pathways.* Bioinformatics, 2000. **16**(8): p. 727-734.
53. Shmulevich, I., et al., *Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks.* Bioinformatics, 2002. **18**(2): p. 261-274.
54. D'haeseleer, P., et al. *Linear modeling of mRNA expression levels during CNS development and injury*. in *Pacific symposium on biocomputing*. 1999.
55. Brazhnik, P., *Inferring gene networks from steady-state response to single-gene perturbations.* Journal of theoretical biology, 2005. **237**(4): p. 427-440.
56. Gardner, T.S., et al., *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. **301**(5629): p. 102-105.
57. Bongard, J. and H. Lipson, *Automated reverse engineering of nonlinear dynamical systems.* Proceedings of the National Academy of Sciences, 2007. **104**(24): p. 9943-9948.
58. Wahde, M. and J. Hertz, *Coarse-grained reverse engineering of genetic regulatory networks.* Biosystems, 2000. **55**(1): p. 129-136.
59. Yip, K.Y., et al., *Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data.* PloS one, 2010. **5**(1): p. e8121.
60. Stark, J., et al., *Reconstructing gene networks: what are the limits?* Biochemical Society Transactions, 2003. **31**(6): p. 1519-1525.
61. Blake, W.J., et al., *Noise in eukaryotic gene expression.* Nature, 2003. **422**(6932): p. 633-637.
62. Thattai, M. and A. Van Oudenaarden, *Intrinsic noise in gene regulatory networks.* Proceedings of the National Academy of Sciences, 2001. **98**(15): p. 8614-8619.
63. Friedman, N., *Inferring cellular networks using probabilistic graphical models.* Science, 2004. **303**(5659): p. 799-805.
64. Cooper, G.F. and E. Herskovits, *A Bayesian method for the induction of probabilistic networks from data.* Machine learning, 1992. **9**(4): p. 309-347.

65. Heckerman, D., *A tutorial on learning with Bayesian networks*, in *Innovations in Bayesian Networks*. 2008, Springer. p. 33-82.
66. TCGAPortal, *Nationl Human Genome Research Institute*. https://tcga-data.nci.nih.gov/tcga/.
67. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
68. Dennis Jr, G., et al., *DAVID: database for annotation, visualization, and integrated discovery.* Genome biol, 2003. **4**(5): p. P3.
69. Kanehisa, M., *The KEGG database.* silico simulation of biological processes, 2002. **247**: p. 91-103.
70. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-1740.
71. Matys, V., et al., *TRANSFAC®: transcriptional regulation, from patterns to profiles.* Nucleic acids research, 2003. **31**(1): p. 374-378.
72. Hsu, S.-D., et al., *miRTarBase: a database curates experimentally validated microRNA–target interactions.* Nucleic acids research, 2010: p. gkq1107.
73. Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* Rna, 2006. **12**(2): p. 192-197.
74. Xiao, F., et al., *miRecords: an integrated resource for microRNA–target interactions.* Nucleic acids research, 2009. **37**(suppl 1): p. D105-D110.
75. Sengupta, D. and S. Bandyopadhyay, *Participation of microRNAs in human interactome: extraction of microRNA–microRNA regulations.* Molecular Biosystems, 2011. **7**(6): p. 1966-1973.
76. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2009. **37**(1): p. 1-13.
77. Zhang, B., S. Kirov, and J. Snoddy, *WebGestalt: an integrated system for exploring gene sets in various biological contexts.* Nucleic Acids Research, 2005. **33**(suppl 2): p. W741-W748.
78. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995: p. 289-300.
79. Zeller, C., et al., *Candidate DNA methylation drivers of acquired cisplatin resistance in ovarian cancer identified by methylome and expression profiling.* Oncogene, 2012. **31**(42): p. 4567-4576.
80. Wolfe, C.J., I.S. Kohane, and A.J. Butte, *Systematic survey reveals general applicability of.* BMC bioinformatics, 2005. **6**(1): p. 227.
81. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC bioinformatics, 2008. **9**(1): p. 559.
82. Jiang, C., et al., *TRED: a transcriptional regulatory element database, new entries and other development.* Nucleic acids research, 2007. **35**(suppl 1): p. D137-D140.
83. Sandelin, A., et al., *JASPAR: an open‐access database for eukaryotic transcription factor binding profiles.* Nucleic acids research, 2004. **32**(suppl 1): p. D91-D94.
84. Marschall, T. and S. Rahmann, *Efficient exact motif discovery.* Bioinformatics, 2009. **25**(12): p. i356-i364.
85. Smoot, M.E., et al., *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 2011. **27**(3): p. 431-432.

86. Hu, Z., et al., *VisANT: an online visualization and analysis tool for biological interaction data.* BMC bioinformatics, 2004. **5**(1): p. 17.
87. Carvalho, A.M., *Scoring functions for learning bayesian networks.* Inesc-id Tec. Rep, 2009.
88. Akulenko, R. and V. Helms, *DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples.* Human molecular genetics, 2013. **22**(15): p. 3016-3022.
89. Dreos, R., et al., *EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era.* Nucleic acids research, 2013. **41**(D1): p. D157-D164.
90. Chu, G., et al., *Significance Analysis of Microarrays Users Guide and Technical Document.* 2001.
91. Hahne, F., et al., *Bioconductor case studies*. 2010: Springer.
92. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, Complex Systems, 2006. **1695**(5).
93. Laczny, C., et al., *miRTrail-a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases.* BMC bioinformatics, 2012. **13**(1): p. 36.
94. Makhorin, A., *GLPK (GNU linear programming kit)*. 2008.
95. Kroshko, D., *OpenOpt.* Software package downloadable from http://openopt. org, 2007.
96. Lu, M., et al., *TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs.* BMC bioinformatics, 2010. **11**(1): p. 419.
97. Hewett, M., et al., *PharmGKB: the pharmacogenetics knowledge base.* Nucleic acids research, 2002. **30**(1): p. 163-165.
98. Davis, A., et al., *CTD-Comparative Toxicogenomics Database.*
99. Ahmed, J., et al., *CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge.* Nucleic acids research, 2011. **39**(suppl 1): p. D960-D967.
100. Carter, H., et al., *Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.* Cancer research, 2009. **69**(16): p. 6660-6667.
101. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-158.
102. Jones, S., et al., *Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.* science, 2008. **321**(5897): p. 1801-1806.
103. Kaminker, J.S., et al., *Distinguishing cancer-associated missense mutations from common polymorphisms.* Cancer research, 2007. **67**(2): p. 465-473.
104. Parsons, D.W., et al., *An integrated genomic analysis of human glioblastoma multiforme.* Science, 2008. **321**(5897): p. 1807-1812.
105. Sjöblom, T., et al., *The consensus coding sequences of human breast and colorectal cancers.* science, 2006. **314**(5797): p. 268-274.
106. Wood, L.D., et al., *The genomic landscapes of human breast and colorectal cancers.* Science, 2007. **318**(5853): p. 1108-1113.
107. Torkamani, A. and N.J. Schork, *Prediction of cancer driver mutations in protein kinases.* Cancer research, 2008. **68**(6): p. 1675-1682.
108. Barnholtz-Sloan, J., et al., *Somatic alterations in brain tumors.* Oncology reports, 2008. **20**(1): p. 203-210.
109. Ng, P.C. and S. Henikoff, *Predicting deleterious amino acid substitutions.* Genome research, 2001. **11**(5): p. 863-874.

110. Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data.* Nucleic acids research, 2011. **39**(suppl 1): p. D152-D157.
111. Keller, A., et al., *Toward the blood-borne miRNome of human diseases.* nature methods, 2011. **8**(10): p. 841-843.
112. Fatemi, M., et al., *Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level.* Nucleic acids research, 2005. **33**(20): p. e176-e176.
113. Gu, Z., et al., *circlize implements and enhances circular visualization in R.* Bioinformatics, 2014: p. btu393.
114. Hamed, M., et al., *TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks.* Nucleic Acids Research, 2015.
115. Yan, Z., et al., *Integrative analysis of gene and miRNA expression profiles with transcription factor–miRNA feed-forward loops identifies regulators in human cancers.* Nucleic acids research, 2012: p. gks395.
116. Li, K., et al., *Functional analysis of microRNA and transcription factor synergistic regulatory network based on identifying regulatory motifs in non-small cell lung cancer.* BMC systems biology, 2013. **7**(1): p. 122.
117. Poos, K., et al., *How microRNA and transcription factor co-regulatory networks affect osteosarcoma cell proliferation.* PLoS computational biology, 2013. **9**(8): p. e1003210.
118. Qin, S., F. Ma, and L. Chen, *Gene regulatory networks by transcription factors and microRNAs in breast cancer.* Bioinformatics, 2014: p. btu597.
119. Griffith, O.L., et al., *ORegAnno: an open-access community-driven resource for regulatory annotation.* Nucleic acids research, 2008. **36**(suppl 1): p. D107-D113.
120. van Rooij, E., et al., *Control of stress-dependent cardiac growth and gene expression by a microRNA.* Science, 2007. **316**(5824): p. 575-579.
121. van Rooij, E., et al., *A family of microRNAs encoded by myosin genes governs myosin expression and muscle performance.* Developmental cell, 2009. **17**(5): p. 662-673.
122. Zisoulis, D.G., et al., *Autoregulation of microRNA biogenesis by let-7 and Argonaute.* Nature, 2012. **486**(7404): p. 541-544.
123. Matkovich, S.J., Y. Hu, and G.W. Dorn, *Regulation of cardiac microRNAs by cardiac microRNAs.* Circulation research, 2013. **113**(1): p. 62-71.
124. Tang, R., et al., *Mouse miRNA-709 directly regulates miRNA-15a/16-1 biogenesis at the posttranscriptional level in the nucleus: evidence for a microRNA hierarchy system.* Cell research, 2012. **22**(3): p. 504-515.
125. Qiu, C., et al., *microRNA evolution in a human transcription factor and microRNA regulatory network.* BMC systems biology, 2010. **4**(1): p. 90.
126. Yang, J.-H., et al., *starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data.* Nucleic acids research, 2011. **39**(suppl 1): p. D202-D209.
127. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-233.
128. Krek, A., et al., *Combinatorial microRNA target predictions.* Nature genetics, 2005. **37**(5): p. 495-500.

129. Miranda, K.C., et al., *A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes.* Cell, 2006. **126**(6): p. 1203-1217.
130. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nature genetics, 2007. **39**(10): p. 1278-1284.
131. John, B., et al., *Human microRNA targets.* PLoS biology, 2004. **2**(11): p. e363.
132. Yang, J.-H., et al., *ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data.* Nucleic acids research, 2013. **41**(D1): p. D177-D187.
133. Lu, M., et al., *An analysis of human microRNA and disease associations.* PloS one, 2008. **3**(10): p. e3420.
134. Queralt-Rosinach, N. and L.I. Furlong. *DisGeNET RDF: A Gene-Disease Association Linked Open Data Resource*. in *SWAT4LS*. 2013.
135. Da Wei Huang, B.T.S. and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2008. **4**(1): p. 44-57.
136. Lopes, C.T., et al., *Cytoscape Web: an interactive web-based network browser.* Bioinformatics, 2010. **26**(18): p. 2347-2348.
137. Rai, M., S. Verma, and S. Tapaswi, *A power aware minimum connected dominating set for wireless sensor networks.* Journal of networks, 2009. **4**(6): p. 511-519.
138. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nature genetics, 2002. **31**(1): p. 64-68.
139. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proceedings of the National Academy of Sciences, 2003. **100**(21): p. 11980-11985.
140. Tsang, J., J. Zhu, and A. van Oudenaarden, *MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals.* Molecular cell, 2007. **26**(5): p. 753-767.
141. O'Donnell, K.A., et al., *c-Myc-regulated microRNAs modulate E2F1 expression.* nature, 2005. **435**(7043): p. 839-843.
142. He, L., et al., *A microRNA component of the p53 tumour suppressor network.* Nature, 2007. **447**(7148): p. 1130-1134.
143. Li, X., et al., *A microRNA imparts robustness against environmental fluctuation during development.* Cell, 2009. **137**(2): p. 273-282.
144. Shalgi, R., et al., *Coupling transcriptional and post-transcriptional miRNA regulation in the control of cell fate.* Aging (Albany NY), 2009. **1**(9): p. 762.
145. El Baroudi, M., et al., *A curated database of miRNA mediated feed-forward loops involving MYC as master regulator.* PloS one, 2011. **6**(3): p. e14742.
146. Friard, O., et al., *CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse.* BMC bioinformatics, 2010. **11**(1): p. 435.
147. Yu, G., et al., *GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.* Bioinformatics, 2010. **26**(7): p. 976-978.
148. Network, C.G.A., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
149. Santarius, T., et al., *A census of amplified and overexpressed human cancer genes.* Nature Reviews Cancer, 2010. **10**(1): p. 59-64.
150. Humbert, P.O., et al., *E2f3 is critical for normal cellular proliferation.* Genes & development, 2000. **14**(6): p. 690-703.

151. Reyes, A., *The Role of E2F3 in the Macrophage Assisted Metastasis of Breast Cancer.* 2007.

152. Vimala, K., et al., *Curtailing Overexpression of E2F3 in Breast Cancer Using< i> siRNA</i>(E2F3)-Based Gene Silencing.* Archives of medical research, 2012. **43**(6): p. 415-422.

153. Nakajima, G., et al., *Non-coding MicroRNAs hsa-let-7g and hsa-miR-181b are Associated with Chemoresponse to S-1 in Colon Cancer.* Cancer Genomics-Proteomics, 2006. **3**(5): p. 317-324.

154. Della Vittoria Scarpati, G., et al., *Analysis of Differential miRNA Expression in Primary Tumor and Stroma of Colorectal Cancer Patients.* BioMed research international, 2014. **2014**.

155. Cheng, H.H., et al., *Circulating microRNA profiling identifies a subset of metastatic prostate cancer patients with evidence of cancer-associated hypoxia.* PloS one, 2013. **8**(7): p. e69239.

156. Garofalo, M. and C.M. Croce, *MicroRNAs as therapeutic targets in chemoresistance.* Drug Resistance Updates, 2013. **16**(3): p. 47-59.

157. Guo, Z.S., et al., *The enhanced tumor selectivity of an oncolytic vaccinia lacking the host range and antiapoptosis genes SPI-1 and SPI-2.* Cancer research, 2005. **65**(21): p. 9991-9998.

158. Rimmelé, P., et al., *Spi-1/PU. 1 oncogene accelerates DNA replication fork elongation and promotes genetic instability in the absence of DNA breakage.* Cancer research, 2010. **70**(17): p. 6757-6766.

159. Kossenkov, A.V., et al., *Resection of non–small cell lung cancers reverses tumor-induced gene expression changes in the peripheral immune system.* Clinical Cancer Research, 2011. **17**(18): p. 5867-5877.

160. Scheiber, M.N., et al., *FLI1 Expression is Correlated with Breast Cancer Cellular Growth, Migration, and Invasion and Altered Gene Expression.* Neoplasia, 2014. **16**(10): p. 801-813.

161. Song, W., et al., *Oncogenic Fli-1 is a potential prognostic marker for the progression of epithelial ovarian cancer.* BMC cancer, 2014. **14**(1): p. 424.

162. Sakurai, T., et al., *Functional roles of Fli‐1, a member of the Ets family of transcription factors, in human breast malignancy.* Cancer science, 2007. **98**(11): p. 1775-1784.

163. Bertucci, F., et al., *Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters.* Oncogene, 2004. **23**(7): p. 1377-1391.

164. Chang, J.C., et al., *Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer.* The Lancet, 2003. **362**(9381): p. 362-369.

165. Sgroi, D.C., et al., *In vivo gene expression profile analysis of human breast cancer progression.* Cancer research, 1999. **59**(22): p. 5656-5661.

166. Birkenkamp-Demtroder, K., et al., *Gene expression in colorectal cancer.* Cancer Research, 2002. **62**(15): p. 4352-4363.

167. Ma, X.-J., et al., *Gene expression profiles of human breast cancer progression.* Proceedings of the National Academy of Sciences, 2003. **100**(10): p. 5974-5979.

168. Yang, L., N. Belaguli, and D.H. Berger, *MicroRNA and colorectal cancer.* World journal of surgery, 2009. **33**(4): p. 638-646.

169. Xi, Y., et al., *Prognostic values of microRNAs in colorectal cancer.* Biomarker insights, 2006. **1**: p. 113.

170. Saito, M., et al., *The association of microRNA expression with prognosis and progression in early-stage, non–small cell lung adenocarcinoma: a retrospective analysis of three cohorts.* Clinical cancer research, 2011. **17**(7): p. 1875-1882.
171. Jiang, Q., et al., *miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic acids research, 2009. **37**(suppl 1): p. D98-D104.
172. Sengupta, D. and S. Bandyopadhyay, *Topological patterns in microRNA–gene regulatory network: studies in colorectal and breast cancer.* Mol. BioSyst., 2013. **9**(6): p. 1360-1371.
173. Ihaka, R. and R. Gentleman, *R: a language for data analysis and graphics.* Journal of computational and graphical statistics, 1996. **5**(3): p. 299-314.
174. Andrews, S., *FastQC: A quality control tool for high throughput sequence data.* Reference Source, 2010.
175. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics (Oxford, England), 2009. **25**: p. 1754-60.
176. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
177. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome research, 2012. **22**: p. 568-76.
178. Gouy, M., S. Guindon, and O. Gascuel, *SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.* Molecular biology and evolution, 2010. **27**: p. 221-4.
179. Reik, W. and J. Walter, *Genomic imprinting: parental influence on the genome.* Nature Reviews Genetics, 2001. **2**(1): p. 21-32.
180. Morison, I.M., J.P. Ramsay, and H.G. Spencer, *A census of mammalian imprinting.* Trends in Genetics, 2005. **21**(8): p. 457-465.
181. Lau, M., et al., *Loss of the imprinted IGF2/cation-independent mannose 6-phosphate receptor results in fetal overgrowth and perinatal lethality.* Genes & development, 1994. **8**(24): p. 2953-2963.
182. HAIG, D. and M. WESTOBY, *Selective forces in the emergence of the seed habit.* Biological Journal of the Linnean Society, 1989. **38**(3): p. 215-238.
183. Moore, T. and D. Haig, *Genomic imprinting in mammalian development: a parental tug-of-war.* Trends in Genetics, 1991. **7**(2): p. 45-49.
184. Barbara, H., et al., *Imprinted genes show unique patterns of sequence conservation.* BMC Genomics. **11**.
185. Morison, I.M., C.J. Paton, and S.D. Cleverley, *The imprinted gene and parent-of-origin effect database.* Nucleic Acids Research, 2001. **29**(1): p. 275-276.
186. Wingender, E., *The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.* Briefings in Bioinformatics, 2008. **9**(4): p. 326-332.
187. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545.
188. Ihaka, R. and R. Gentleman, *R: A language for data analysis and graphics.* Journal of computational and graphical statistics, 1996: p. 299-314.
189. Fitzpatrick, G.V., P.D. Soloway, and M.J. Higgins, *Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1.* Nature genetics, 2002. **32**(3): p. 426-431.

190. Brown, K.W., et al., *Imprinting mutation in the Beckwith-Wiedemann syndrome leads to biallelic IGF2 expression through an H19-independent pathway.* Human molecular genetics, 1996. **5**(12): p. 2027-2032.

191. Jerome, C., et al., *Assignment of growth factor receptor-bound protein 10 (GRB10) to human chromosome 7p11. 2-p12.* Genomics, 1997. **40**(1): p. 215-216.

192. Mori, K., B. Giovannone, and R.J. Smith, *Distinct Grb10 domain requirements for effects on glucose uptake and insulin signaling.* Molecular and cellular endocrinology, 2005. **230**(1): p. 39-50.

193. Tycko, B. and I.M. Morison, *Physiological functions of imprinted genes.* Journal of cellular physiology, 2002. **192**(3): p. 245-258.

194. Kent, L., et al., *Beckwith Weidemann syndrome: A behavioral phenotype–genotype study.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2008. **147B**(7): p. 1295-1297.

195. Steinhoff, C., et al., *Expression profile and transcription factor binding site exploration of imprinted genes in human and mouse.* BMC Genomics, 2009. **10**(1): p. 144.

196. Krüger, I., et al., *Sp1/Sp3 compound heterozygous mice are not viable: impaired erythropoiesis and severe placental defects.* Developmental Dynamics, 2007. **236**(8): p. 2235-2244.

197. Varrault, A., et al., *Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth.* Developmental cell, 2006. **11**(5): p. 711-722.

198. Berg, J.S., et al., *Imprinted Genes That Regulate Early Mammalian Growth Are Coexpressed in Somatic Stem Cells.* PloS one, 2011. **6**(10): p. e26410.

199. Bain, G., et al., *Both E12 and E47 allow commitment to the B cell lineage.* Immunity, 1997. **6**(2): p. 145-154.

200. Thomas, K., et al., *SP1 transcription factors in male germ cell development and differentiation.* Molecular and cellular endocrinology, 2007. **270**(1-2): p. 1-7.

201. Moignard, V., et al., *Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis.* Nature cell biology, 2013. **15**(4): p. 363-372.

202. Som, A., et al., *The PluriNetWork: An electronic representation of the network underlying pluripotency in mouse, and its applications.* PloS one, 2010. **5**(12): p. e15165.

203. Park, S.-J., et al., *Computational Promoter Modeling Identifies the Modes of Transcriptional Regulation in Hematopoietic Stem Cells.* PloS one, 2014. **9**(4): p. e93853.

204. Cabezas-Wallscheid, N., et al., *Identification of Regulatory Networks in HSCs and Their Immediate Progeny via Integrated Proteome, Transcriptome, and DNA Methylome Analysis.* Cell stem cell, 2014. **15**(4): p. 507-522.

205. Klimmeck, D., et al., *Transcriptome-wide Profiling and Posttranscriptional Analysis of Hematopoietic Stem/Progenitor Cell Differentiation toward Myeloid Commitment.* Stem cell reports, 2014. **3**(5): p. 858-875.

206. Abadie, C., et al., *Acute lymphocytic leukaemia in a child with Beckwith-Wiedemann syndrome harbouring a CDKN1C mutation.* Eur J Med Genet, 2010. **53**(6): p. 400-3.

207. Venkatraman, A., et al., *Maternal imprinting at the H19-Igf2 locus maintains adult haematopoietic stem cell quiescence.* Nature, 2013.

208. Hamed, M., et al., *Cellular functions of genetically imprinted genes in human and mouse as annotated in the gene ontology.* PLoS One, 2012. **7**(11): p. e50285.

209. Scandura, J.M., et al., *Transforming growth factor β-induced cell cycle arrest of human hematopoietic cells requires p57KIP2 up-regulation.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(42): p. 15231-15236.
210. Kwon, K., et al., *Instructive role of the transcription factor E2A in early B lymphopoiesis and germinal center B cell development.* Immunity, 2008. **28**(6): p. 751-762.
211. Venkatraman, A., et al., *Maternal imprinting at the H19-Igf2 locus maintains adult haematopoietic stem cell quiescence.* Nature, 2013.
212. Hutter, B., et al., *Imprinted genes show unique patterns of sequence conservation.* BMC genomics, 2010. **11**(1): p. 649.
213. Hutter, B., et al., *Divergence of imprinted genes during mammalian evolution.* BMC evolutionary biology, 2010. **10**(1): p. 116.
214. Schulz, R., et al., *WAMIDEX: a web atlas of murine genomic imprinting and differential expression.* Epigenetics: official journal of the DNA Methylation Society, 2008. **3**(2): p. 89.
215. Chambers, S.M., et al., *Hematopoietic fingerprints: an expression database of stem cells and their progeny.* Cell Stem Cell, 2007. **1**(5): p. 578-591.
216. Di Tullio, A., et al., *CCAAT/enhancer binding protein α (C/EBPα)-induced transdifferentiation of pre-B cells into macrophages involves no overt retrodifferentiation.* Proceedings of the National Academy of Sciences, 2011. **108**(41): p. 17016-17021.
217. Seita, J., et al., *Gene Expression Commons: An Open Platform for Absolute Gene Expression Profiling.* PloS one, 2012. **7**(7): p. e40321.
218. Lattin, J.E., et al., *Expression analysis of G Protein-Coupled Receptors in mouse macrophages.* Immunome research, 2008. **4**(1): p. 5.
219. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Research, 2002. **30**(1): p. 207-210.
220. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome biology, 2004. **5**(10): p. R80.
221. Hahne, F. and R. Gentleman, *Bioconductor case studies.* 2008: Springer.
222. Schlicker, A. and M. Albrecht, *FunSimMat: a comprehensive functional similarity database.* Nucleic Acids Research, 2008. **36**(suppl 1): p. D434-D439.
223. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms.* PloS one, 2011. **6**(7): p. e21800.
224. Grumont, R. and S. Gerondakis, *The murine c-rel proto-oncogene encodes two mRNAs the expression of which is modulated by lymphoid stimuli.* Oncogene research, 1990. **5**(4): p. 245.
225. Bult, C., et al., *P4-S The Mouse Genome Informatics Database: An Integrated Resource for Mouse Genetics and Genomics.* Journal of Biomolecular Techniques: JBT, 2007. **18**(1): p. 2.
226. Ravasz, E., et al., *Hierarchical Organization of Modularity in Metabolic Networks.* Science, 2002. **297**(5586): p. 1551-1555.
227. Göke, J., et al., *Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development.* PLoS computational biology, 2011. **7**(12): p. e1002304.
228. Sun, J.C., J.N. Beilke, and L.L. Lanier, *Adaptive immune features of natural killer cells.* Nature, 2009. **457**(7229): p. 557-561.

229. Haug, J.S., et al., *N-cadherin expression level distinguishes reserved versus primed states of hematopoietic stem cells.* Cell stem cell, 2008. **2**(4): p. 367-379.
230. Månsson, R., et al., *Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors.* Immunity, 2007. **26**(4): p. 407-419.
231. Gekas, C. and T. Graf, *CD41 expression marks myeloid-biased adult hematopoietic stem cells and increases with age.* Blood, 2013. **121**(22): p. 4463-4472.
232. Sanjuan-Pla, A., et al., *Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy.* Nature, 2013. **502**(7470): p. 232-236.
233. Yamanaka, S., *Induction of pluripotent stem cells from mouse fibroblasts by four transcription factors.* Cell proliferation, 2008. **41**: p. 51-56.
234. Takahashi, K., et al., *Induction of pluripotent stem cells from fibroblast cultures.* Nature protocols, 2007. **2**(12): p. 3081-3089.
235. Gearhart, J., E.E. Pashos, and M.K. Prasad, *Pluripotency redux--advances in stem-cell research.* N Engl J Med, 2007. **357**(15): p. 1469-72.
236. de Alboran, I.M., et al., *Analysis of C-MYC function in normal cells via conditional gene-targeted mutation.* Immunity, 2001. **14**(1): p. 45-55.
237. Bluteau, O., et al., *Developmental changes in human megakaryopoiesis.* Journal of Thrombosis and Haemostasis, 2013.
238. Pang, C.J., et al., *Kruppel-like factor 1 (KLF1), KLF2, and Myc control a regulatory network essential for embryonic erythropoiesis.* Mol Cell Biol, 2012. **32**(13): p. 2628-44.
239. Capron, C., et al., *A major role of TGF-beta1 in the homing capacities of murine hematopoietic stem cell/progenitors.* Blood, 2010. **116**(8): p. 1244-53.
240. Wang, D., I. Paz-Priel, and A.D. Friedman, *NF-kappa B p50 regulates C/EBP alpha expression and inflammatory cytokine-induced neutrophil production.* J Immunol, 2009. **182**(9): p. 5757-62.
241. Cokic, V.P., et al., *JAK-STAT and AKT pathway-coupled genes in erythroid progenitor cells through ontogeny.* J Transl Med, 2012. **10**: p. 116.
242. Resendes, K.K. and A.G. Rosmarin, *Sp1 control of gene expression in myeloid cells.* Crit Rev Eukaryot Gene Expr, 2004. **14**(3): p. 171-81.
243. Kruger, I., et al., *Sp1/Sp3 compound heterozygous mice are not viable: impaired erythropoiesis and severe placental defects.* Dev Dyn, 2007. **236**(8): p. 2235-44.
244. Macaluso, M., M. Montanari, and A. Giordano, *The regulation of ER-α transcription by pRb2/p130 in breast cancer.* Annals of Oncology, 2005. **16**(suppl 4): p. iv20-iv22.
245. Siegel, R., et al., *Cancer statistics, 2014.* CA: a cancer journal for clinicians, 2014. **64**(1): p. 9-29.
246. Volinia, S. and C.M. Croce, *Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer.* Proceedings of the National Academy of Sciences, 2013. **110**(18): p. 7413-7417.
247. Cava, C., et al., *Integration of mRNA Expression Profile, Copy Number Alterations, and microRNA Expression Levels in Breast Cancer to Improve Grade Definition.* PloS one, 2014. **9**(5): p. e97681.
248. West, J., et al., *Differential network entropy reveals cancer system hallmarks.* Scientific reports, 2012. **2**.
249. Teschendorff, A.E. and S. Severini, *Increased entropy of signal transduction in the cancer metastasis phenotype.* BMC systems biology, 2010. **4**(1): p. 104.

250. Schramm, G., N. Kannabiran, and R. König, *Regulation patterns in signaling networks of cancer.* BMC systems biology, 2010. **4**(1): p. 162.
251. Tuck, D.P., H.M. Kluger, and Y. Kluger, *Characterizing disease states from topological properties of transcriptional regulatory networks.* BMC bioinformatics, 2006. **7**(1): p. 236.
252. Pujana, M.A., et al., *Network modeling links breast cancer susceptibility and centrosome dysfunction.* Nature genetics, 2007. **39**(11): p. 1338-1349.
253. Platzer, A., et al., *Characterization of protein-interaction networks in tumors.* BMC bioinformatics, 2007. **8**(1): p. 224.
254. Ulitsky, I. and R. Shamir, *Identification of functional modules using network topology and high-throughput data.* BMC systems biology, 2007. **1**(1): p. 8.
255. Chuang, H.Y., et al., *Network ‐ based classification of breast cancer metastasis.* Molecular systems biology, 2007. **3**(1).
256. Milanesi, L., et al., *Trends in modeling biomedical complex systems.* BMC bioinformatics, 2009. **10**(Suppl 12): p. I1.
257. Taylor, I.W., et al., *Dynamic modularity in protein interaction networks predicts breast cancer outcome.* Nature biotechnology, 2009. **27**(2): p. 199-204.
258. Hudson, N.J., A. Reverter, and B.P. Dalrymple, *A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation.* PLoS computational biology, 2009. **5**(5): p. e1000382.
259. Nibbe, R.K., M. Koyutürk, and M.R. Chance, *An integrative-omics approach to identify functional sub-networks in human colorectal cancer.* PLoS computational biology, 2010. **6**(1): p. e1000639.
260. Yao, C., et al., *Multi-level reproducibility of signature hubs in human interactome for breast cancer metastasis.* BMC systems biology, 2010. **4**(1): p. 151.
261. Komurov, K., M.A. White, and P.T. Ram, *Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data.* PLoS computational biology, 2010. **6**(8): p. e1000889.
262. Komurov, K. and P.T. Ram, *Patterns of human gene expression variance show strong associations with signaling network hierarchy.* BMC systems biology, 2010. **4**(1): p. 154.
263. Alzate, O. and A. Vazquez, *Protein Interaction Networks.* 2010.
264. Olex, A.L., et al., *Integration of gene expression data with network-based analysis to identify signaling and metabolic pathways regulated during the development of osteoarthritis.* Gene, 2014. **542**(1): p. 38-45.
265. Califano, A., *Rewiring makes the difference.* Molecular Systems Biology, 2011. **7**(1).
266. Bandyopadhyay, S., et al., *Rewiring of genetic networks in response to DNA damage.* Science, 2010. **330**(6009): p. 1385-1389.
267. Ideker, T. and N.J. Krogan, *Differential network biology.* Molecular systems biology, 2012. **8**(1).
268. Tesson, B.M., R. Breitling, and R.C. Jansen, *DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules.* BMC bioinformatics, 2010. **11**(1): p. 497.
269. Zhang, B., et al., *DDN: a caBIG® analytical tool for differential network analysis.* Bioinformatics, 2011. **27**(7): p. 1036-1038.
270. Jones, M.E., et al., *Endometrial cancer survival after breast cancer in relation to tamoxifen treatment: pooled results from three countries.* Breast Cancer Res, 2012. **14**(3): p. R91.

271. Gasco, M., S. Shami, and T. Crook, *The p53 pathway in breast cancer.* Breast Cancer Research, 2002. **4**(2): p. 70.

272. Walerych, D., et al., *The rebel angel: mutant p53 as the driving oncogene in breast cancer.* Carcinogenesis, 2012. **33**(11): p. 2007-2017.

273. Lacroix, M., R.-A. Toillon, and G. Leclercq, *p53 and breast cancer, an update.* Endocrine-related cancer, 2006. **13**(2): p. 293-325.

274. Turner, N., et al., *Targeting triple negative breast cancer: Is p53 the answer?* Cancer treatment reviews, 2013. **39**(5): p. 541-550.

275. Scata, K.A. and W.S. El-Deiry, *p53, BRCA1 and breast Cancer chemoresistance*, in *Breast Cancer Chemosensitivity*. 2007, Springer. p. 70-86.

276. Slyper, M., et al., *Control of Breast Cancer Growth and Initiation by the Stem Cell–Associated Transcription Factor TCF3.* Cancer research, 2012. **72**(21): p. 5613-5624.

277. Chhabra, A., et al., *Expression of transcription factor CREB1 in human breast cancer and its correlation with prognosis.* Oncology reports, 2007. **18**(4): p. 953-958.

278. Haakenson, J.K., M. Kester, and D.X. Liu, *The ATF/CREB family of transcription factors in breast cancer.* Targeting New Pathways and Cell Death in Breast Cancer, 2012.

279. Dong, L., et al., *Mechanisms of transcriptional activation of bcl-2gene expression by 17β-estradiol in breast cancer cells.* Journal of Biological Chemistry, 1999. **274**(45): p. 32099-32107.

280. Zhang, S., et al., *ROR1 is expressed in human breast cancer and associated with enhanced tumor-cell growth.* PloS one, 2012. **7**(3): p. e31127.

281. Xiao, X., et al., *Targeting CREB for cancer therapy: friend or foe.* Current cancer drug targets, 2010. **10**(4): p. 384-391.

282. Sakamoto, K.M. and D.A. Frank, *CREB in the pathophysiology of cancer: implications for targeting transcription factors for cancer therapy.* Clinical Cancer Research, 2009. **15**(8): p. 2583-2587.

283. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic acids research, 2010: p. gkq929.

284. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nature methods, 2010. **7**(4): p. 248-249.

285. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic acids research, 2003. **31**(13): p. 3812-3814.

286. Tian, Y., et al. *Knowledge-guided differential dependency network learning for detecting structural changes in biological networks*. in *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. 2011. ACM.

287. Cines, D.B., et al., *Endothelial cells in physiology and in the pathophysiology of vascular disorders.* Blood, 1998. **91**(10): p. 3527-61.

288. Ballard, V.L. and J.M. Edelberg, *Targets for regulating angiogenesis in the ageing endothelium.* Expert Opin Ther Targets, 2007. **11**(11): p. 1385-99.

289. Werner, N. and G. Nickenig, *Clinical and therapeutical implications of EPC biology in atherosclerosis.* J Cell Mol Med, 2006. **10**(2): p. 318-32.

290. Menegazzo, L., et al., *Endothelial progenitor cells in diabetes mellitus.* Biofactors, 2012. **38**(3): p. 194-202.

291. Ingram, D.A., et al., *Vessel wall-derived endothelial cells rapidly proliferate because they contain a complete hierarchy of endothelial progenitor cells.* Blood, 2005. **105**(7): p. 2783-6.
292. Asahara, T., et al., *Isolation of putative progenitor endothelial cells for angiogenesis.* Science, 1997. **275**(5302): p. 964-7.
293. Barthelmes, D., et al., *Isolation and characterization of mouse bone marrow-derived Lin(-)/VEGF-R2(+) progenitor cells.* Ann Hematol, 2013. **92**(11): p. 1461-72.
294. Shi, Q., et al., *Evidence for circulating bone marrow-derived endothelial cells.* Blood, 1998. **92**(2): p. 362-7.
295. Delamaire, M., et al., *Impaired leucocyte functions in diabetic patients.* Diabet Med, 1997. **14**(1): p. 29-34.
296. Geerlings, S.E. and A.I. Hoepelman, *Immune dysfunction in patients with diabetes mellitus (DM).* FEMS Immunol Med Microbiol, 1999. **26**(3-4): p. 259-65.
297. Sato, N. and H. Shimizu, *Granulocyte-colony stimulating factor improves an impaired bactericidal function in neutrophils from STZ-induced diabetic rats.* Diabetes, 1993. **42**(3): p. 470-3.
298. Loomans, C.J., et al., *Endothelial progenitor cell dysfunction: a novel concept in the pathogenesis of vascular complications of type 1 diabetes.* Diabetes, 2004. **53**(1): p. 195-9.
299. Awad, O., et al., *Obese diabetic mouse environment differentially affects primitive and monocytic endothelial cell progenitors.* Stem Cells, 2005. **23**(4): p. 575-83.
300. Tepper, O.M., et al., *Human endothelial progenitor cells from type II diabetics exhibit impaired proliferation, adhesion, and incorporation into vascular structures.* Circulation, 2002. **106**(22): p. 2781-6.
301. Segal, M.S., et al., *Nitric oxide cytoskeletal-induced alterations reverse the endothelial progenitor cell migratory defect associated with diabetes.* Diabetes, 2006. **55**(1): p. 102-9.
302. Awad, O., et al., *Differential healing activities of CD34+ and CD14+ endothelial cell progenitors.* Arterioscler Thromb Vasc Biol, 2006. **26**(4): p. 758-64.
303. Schatteman, G.C., *Adult bone marrow-derived hemangioblasts, endothelial cell progenitors, and EPCs.* Curr Top Dev Biol, 2004. **64**: p. 141-80.
304. Fadini, G.P., et al., *Significance of endothelial progenitor cells in subjects with diabetes.* Diabetes Care, 2007. **30**(5): p. 1305-13.
305. Barthelmes, D., et al., *Diabetes impairs mobilization of mouse bone marrow-derived Lin(-)/VEGF-R2(+) progenitor cells.* Blood Cells Mol Dis, 2013. **51**(3): p. 163-73.
306. Ferraro, F., et al., *Diabetes impairs hematopoietic stem cell mobilization by altering niche function.* Sci Transl Med, 2011. **3**(104): p. 104ra101.
307. Mukai, N., et al., *A comparison of the tube forming potentials of early and late endothelial progenitor cells.* Exp Cell Res, 2008. **314**(3): p. 430-40.
308. Barthelmes, D., Irhimeh, M.R., Gillies, M.C., Karimipour, M., Zhou, M., Zhu L., Shen, W.Y., *Diabetes impairs mobilization of mouse bone marrow-derived Lin$^-$/VEGF-R2$^+$ progenitor cells.* Blood Cells Mol. Diseases, 2013. http://dx.doi.org/10.1016/j.bcmd.2013.05.002: p. In press.
309. Akulenko, R. and V. Helms, *DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples.* Hum Mol Genet, 2013. **22**(15): p. 3016-22.

310. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-1548.

311. Kim, K.A., et al., *Dysfunction of endothelial progenitor cells under diabetic conditions and its underlying mechanisms.* Arch Pharm Res, 2012. **35**(2): p. 223-34.

312. Barthelmes, D., et al., *Differential gene expression in Lin-/VEGF-R2+ bone marrow-derived endothelial progenitor cells isolated from diabetic mice.* Cardiovasc Diabetol, 2014. **13**(1): p. 42.

313. von Heydebreck, A., B. Gunawan, and L. Fuzesi, *Maximum likelihood estimation of oncogenetic tree models.* Biostatistics, 2004. **5**(4): p. 545-56.

314. Park, B.H., B. Vogelstein, and K.W. Kinzler, *Genetic disruption of PPARdelta decreases the tumorigenicity of human colon cancer cells.* Proc Natl Acad Sci U S A, 2001. **98**(5): p. 2598-603.

315. Wang, Z., et al., *PPARalpha Regulates Mobilization and Homing of Endothelial Progenitor Cells through the HIF-1/SDF-1 Pathway.* Invest Ophthalmol Vis Sci, 2014.

316. Nakae, J., et al., *The forkhead transcription factor Foxo1 regulates adipocyte differentiation.* Dev Cell, 2003. **4**(1): p. 119-29.

317. Betts, D.H. and P. Madan, *Permanent embryo arrest: molecular and cellular concepts.* Mol Hum Reprod, 2008. **14**(8): p. 445-53.

318. Di Stefano, V., et al., *p66ShcA modulates oxidative stress and survival of endothelial progenitor cells in response to high glucose.* Cardiovasc Res, 2009. **82**(3): p. 421-9.

319. Kusuyama, T., et al., *Effects of treatment for diabetes mellitus on circulating vascular progenitor cells.* J Pharmacol Sci, 2006. **102**(1): p. 96-102.

320. van Ark, J., et al., *Type 2 diabetes mellitus is associated with an imbalance in circulating endothelial and smooth muscle progenitor cell numbers.* Diabetologia, 2012. **55**(9): p. 2501-12.

321. Ingram, D.A., et al., *In vitro hyperglycemia or a diabetic intrauterine environment reduces neonatal endothelial colony-forming cell numbers and function.* Diabetes, 2008. **57**(3): p. 724-31.

322. Aicher, A., et al., *Essential role of endothelial nitric oxide synthase for mobilization of stem and progenitor cells.* Nat Med, 2003. **9**(11): p. 1370-6.

323. Oyadomari, S., et al., *Coinduction of endothelial nitric oxide synthase and arginine recycling enzymes in aorta of diabetic rats.* Nitric Oxide, 2001. **5**(3): p. 252-60.

324. Dimmeler, S., E. Dernbach, and A.M. Zeiher, *Phosphorylation of the endothelial nitric oxide synthase at ser-1177 is required for VEGF-induced endothelial cell migration.* FEBS Lett, 2000. **477**(3): p. 258-62.

325. Urbich, C., et al., *Dephosphorylation of endothelial nitric oxide synthase contributes to the anti-angiogenic effects of endostatin.* FASEB J, 2002. **16**(7): p. 706-8.

326. Kaur, S., et al., *Genetic engineering with endothelial nitric oxide synthase improves functional properties of endothelial progenitor cells from patients with coronary artery disease: an in vitro study.* Basic Res Cardiol, 2009. **104**(6): p. 739-49.

327. Hur, J., et al., *Characterization of two types of endothelial progenitor cells and their different contributions to neovasculogenesis.* Arterioscler Thromb Vasc Biol, 2004. **24**(2): p. 288-93.

328. Fadini, G.P., et al., *Circulating endothelial progenitor cells are reduced in peripheral vascular complications of type 2 diabetes mellitus.* J Am Coll Cardiol, 2005. **45**(9): p. 1449-57.

329. Brunner, S., et al., *Correlation of different circulating endothelial progenitor cells to stages of diabetic retinopathy: first in vivo data.* Invest Ophthalmol Vis Sci, 2009. **50**(1): p. 392-8.
330. Yun, H.J. and D.Y. Jo, *Production of stromal cell-derived factor-1 (SDF-1)and expression of CXCR4 in human bone marrow endothelial cells.* J Korean Med Sci, 2003. **18**(5): p. 679-85.
331. De Falco, E., et al., *Altered SDF-1-mediated differentiation of bone marrow-derived endothelial progenitor cells in diabetes mellitus.* J Cell Mol Med, 2009. **13**(9B): p. 3405-14.
332. Janic, B. and A.S. Arbab, *The role and therapeutic potential of endothelial progenitor cells in tumor neovascularization.* ScientificWorldJournal, 2010. **10**: p. 1088-99.
333. Avci-Adali, M., et al., *Porcine EPCs downregulate stem cell markers and upregulate endothelial maturation markers during in vitro cultivation.* J Tissue Eng Regen Med, 2009. **3**(7): p. 512-20.
334. Navarro-Gonzalez, J.F., et al., *Inflammatory molecules and pathways in the pathogenesis of diabetic nephropathy.* Nat Rev Nephrol, 2011. **7**(6): p. 327-40.
335. Cantaluppi, V., et al., *Microvesicles derived from endothelial progenitor cells enhance neoangiogenesis of human pancreatic islets.* Cell Transplant, 2012.
336. Ganasyam, S.R., et al., *Association of Estrogen Receptor-alpha Gene & Metallothionein-1 Gene Polymorphisms in Type 2 Diabetic Women of Andhra Pradesh.* Indian J Clin Biochem, 2012. **27**(1): p. 69-73.
337. Harding, T.C., et al., *Blockade of nonhormonal fibroblast growth factors by FP-1039 inhibits growth of multiple types of cancer.* Sci Transl Med, 2013. **5**(178): p. 178ra39.
338. Thum, T., et al., *Endothelial nitric oxide synthase uncoupling impairs endothelial progenitor cell mobilization and function in diabetes.* Diabetes, 2007. **56**(3): p. 666-74.
339. Kluytmans, J., A. Van Belkum, and H. Verbrugh, *Nasal Carriage of Staphylococcus aureus : Epidemiology , Underlying Mechanisms , and Associated Risks.* Clinical microbiology reviews, 1997. **10**: p. 505-520.
340. Young, B.C., et al., *Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease.* Proceedings of the National Academy of Sciences, 2012. **109**.
341. Köck, R., et al., *Methicillin-resistant Staphylococcus aureus (MRSA): burden of disease and control challenges in Europe.* Euro surveillance : bulletin Européen sur les maladies transmissibles = European communicable disease bulletin, 2010. **15**: p. 19688.
342. Seybold, U., et al., *Emergence of community-associated methicillin-resistant Staphylococcus aureus USA300 genotype as a major cause of health care-associated blood stream infections.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2006. **42**: p. 647-56.
343. Chambers, H.F., *The changing epidemiology of Staphylococcus aureus?* Emerging infectious diseases, 2001. **7**: p. 178-82.
344. Chua, K., et al., *Antimicrobial resistance: Not community-associated methicillin-resistant Staphylococcus aureus (CA-MRSA)! A clinician's guide to community MRSA - its evolving antimicrobial resistance and implications for therapy.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2011. **52**: p. 99-114.

345. Von Eiff, C., et al., *Nasal Carriage as a Source of Staphylococcus aureus Bacteremia.* New England Journal of Medicine, 2001. **344**: p. 11-16.

346. Sabat, A., et al., *Overview of molecular typing methods for outbreak detection and epidemiological surveillance.* Euro surveillance : bulletin Européen sur les maladies transmissibles = European communicable disease bulletin, 2013. **18**: p. 20380.

347. Deurenberg, R.H. and E.E. Stobberingh, *The evolution of Staphylococcus aureus.* Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 2008. **8**: p. 747-63.

348. Nübel, U., et al., *Frequent emergence and limited geographic dispersal of methicillin-resistant Staphylococcus aureus.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**: p. 14130-5.

349. Koreen, L., et al., *spa Typing Method for Discriminating among Staphylococcus aureus Isolates : Implications for Use of a Single Marker To Detect Genetic Micro- and Macrovariation.* Journal of Clinical Microbiology, 2004. **42**.

350. Harmsen, D., et al., *Typing of Methicillin-Resistant Staphylococcus aureus in a University Hospital Setting by Using Novel Software for spa Repeat Determination and Database Management.* Journal of clinical microbiology, 2003. **41**: p. 5442-5448.

351. Witte, W., et al., *Emergence and spread of antibiotic-resistant Gram-positive bacterial pathogens.* International journal of medical microbiology : IJMM, 2008. **298**: p. 365-77.

352. Nübel, U., et al., *Single-nucleotide polymorphism genotyping identifies a locally endemic clone of methicillin-resistant Staphylococcus aureus.* PloS one, 2012. **7**: p. e32698.

353. Engelthaler, D.M., et al., *Rapid and robust phylotyping of spa t003, a dominant MRSA clone in Luxembourg and other European countries.* BMC infectious diseases, 2013. **13**: p. 339.

354. Herrmann, M., et al., *Methicillin-resistant Staphylococcus aureus in Saarland, Germany: a statewide admission prevalence screening study.* PloS one, 2013. **8**: p. e73876.

355. Leopold, S.R., et al., *Bacterial whole genome sequencing revisited: portable, scalable and standardized analysis for typing and detection of virulence and antibiotic resistance genes.* Journal of clinical microbiology, 2014.

356. Harrison, E.M., et al., *Whole genome sequencing identifies zoonotic transmission of MRSA isolates with the novel mecA homologue mecC.* EMBO molecular medicine, 2013. **5**: p. 509-15.

357. Köser, C.U., et al., *Importance of the genetic diversity within the Mycobacterium tuberculosis complex for the development of novel antibiotics and diagnostic tests of drug resistance.* Antimicrobial agents and chemotherapy, 2012. **56**: p. 6080-7.

358. Billal, D.S., et al., *Whole genome analysis of linezolid resistance in Streptococcus pneumoniae reveals resistance and compensatory mutations.* BMC genomics, 2011. **12**: p. 512.

359. Laabei, M., et al., *Predicting the virulence of MRSA from its genome sequence.* Genome research, 2014. **24**: p. 839-49.

360. Fitzgerald, J.R., *Evolution of Staphylococcus aureus during human colonization and infection.* Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 2014. **21**: p. 542-7.

361. Gordon, R.J. and F.D. Lowy, *Pathogenesis of methicillin-resistant Staphylococcus aureus infection.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2008. **46 Suppl 5**: p. S350-9.

362. Wattam, A.R., et al., *PATRIC, the bacterial bioinformatics database and analysis resource.* Nucleic acids research, 2014. **42**: p. D581-91.

363. Chen, L., et al., *VFDB: a reference database for bacterial virulence factors.* Nucleic acids research, 2005. **33**: p. D325-8.

364. Price, J.R., et al., *Whole-Genome Sequencing Shows That Patient-to-Patient Transmission Rarely Accounts for Acquisition of Staphylococcus aureus in an Intensive Care Unit.* Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 2014. **58**: p. 609-18.

365. Mossong, J., et al., *Prevalence, risk factors and molecular epidemiology of methicillin-resistant Staphylococcus aureus (MRSA) colonization in residents of long-term care facilities in Luxembourg, 2010.* Epidemiology and infection, 2013. **141**: p. 1199-206.

366. Ruffing, U., et al., *Matched-cohort DNA microarray diversity analysis of methicillin sensitive and methicillin resistant Staphylococcus aureus isolates from hospital admission patients.* PloS one, 2012. **7**: p. e52487.

367. Huang, D.W., B.T. Sherman, and R.a. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2009. **4**: p. 44-57.

368. Priest, N.K., et al., *From genotype to phenotype: can systems biology be used to predict Staphylococcus aureus virulence?* Nature reviews. Microbiology, 2012. **10**: p. 791-7.

369. Nair, D., et al., *Whole-genome sequencing of Staphylococcus aureus strain RN4220, a key laboratory strain used in virulence research, identifies mutations that affect not only virulence factors but also the fitness of the strain.* Journal of bacteriology, 2011. **193**: p. 2332-5.

370. Chua, K.Y.L., et al., *Population genetics and the evolution of virulence in Staphylococcus aureus.* Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 2014. **21**: p. 554-62.

371. Golubchik, T., et al., *Within-host evolution of Staphylococcus aureus during asymptomatic carriage.* PloS one, 2013. **8**: p. e61319.

372. Lindsay, J.A., *Staphylococcus aureus genomics and the impact of horizontal gene transfer.* International journal of medical microbiology : IJMM, 2014. **304**: p. 103-9.

373. Clarke, S.R., et al., *Analysis of Ebh, a 1.1-megadalton cell wall-associated fibronectin-binding protein of Staphylococcus aureus.* Infection and immunity, 2002. **70**.

374. Highlander, S.K., et al., *Subtle genetic changes enhance virulence of methicillin resistant and sensitive Staphylococcus aureus.* BMC microbiology, 2007. **7**: p. 99.