What we leave behind: reproducibility in chromatin analysis within and across species

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von
Peter Ebert

Saarbrücken

2018

Tag des Kolloquiums: 2019-03-14

Dekan: Prof. Dr. Sebastian Hack

Berichterstatter

Erstgutachter: Prof. Dr. Dr. Thomas Lengauer

Zweitgutachter: Prof. Dr. Hans-Peter Lenhof

Vorsitz: Prof. Dr. Gerhard Weikum

Akad. Mitarbeiter: Dr. Erisa Terolli

Plagiarism note: This document was scanned for plagiarism using the iThenticate service as offered via the Max Planck Society user license: https://ithenticate.mpdl.mpg.de. All figures, tables, and equations were removed from the text prior to the iThenticate scan. Citations in double quotes were kept in the text, but ignored during the scan. For the unpublished work of Chapters 4 and 5, the manuscript preprints (see Appendix E) were excluded from the list of sources. iThenticate's Similarity Index per scanned part of this thesis is given below.

Version of the scanned text:
Draft 2018-12-04 16:31

Access to iThenticate reports:
github.mpi-klsb.mpg.de/pebert/dissertation: see subfolder `plag_check` (MPI login required)

Content	Similarity Index
Abstract and Chapter 1	2%
Chapter 2	1%
Chapter 3	1%
Chapter 4	1%
Chapter 5	1%
Chapter 6	2%

Peter Ebert

Max Planck Institute for Informatics and Saarland University
Saarland Informatics Campus, Germany

“Talk is cheap. Show me the code.”

Linus Torvalds
(linux–kernel mainling list, 2000-08-25)

Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Dr. Dr. Thomas Lengauer and my two advisors Dr. Marcel H. Schulz (chapters 3 and 4) and Dr. Christoph Bock (chapter 5) for their support and scientific input.

I thank all members of the thesis committee for the willingness to invest their time and effort in the doctoral examination procedure.

Of the many people I had the privilege to work with over the years, I can only name a few here, but I am indebted to anyone who offered honest advice, new perspectives, or who let me take a leaf out of their book. Thanks to (in alphabetical order):

Abdulrahman Salhab, Anna Hildebrandt (née Dehof), Anne-Christin Hauschild, Bastian Beggel, Daniel Stöckel, Dilip Durai, Fabian Müller, Fatemeh Behjati, Felipe Albrecht, Florian Schmidt, Gilles Gasparoni, Glenn Lawyer, Jennifer Gerling, Jörn Walter, Karl Nordström, Konstantin Halachev, Lara Schneider, Lars Feuerbach, Lisa Handl, Markus List, Martin Vingron, Matthias Dietzen, Matthias Döring, Michael Scherer, Nico Pfeifer, Nina Gasparoni, Pavlo Lutsik, Sarvesh Nikumbh, Sivarajan Karunanithi, Stefan Bender, Thomas Manke, Tim Kehl, Uwe Brahm, Wolfram Wagner, Yassen Assenov, and all DEEP consortium members.

I am particularly thankful to Ruth Schneppen-Christmann, Joachim Büch and Georg Friedrich for their administrative and technical support throughout the years.

My deepest gratitude for countless insightful, entertaining, and challenging discussions goes to Anna Hake (née Feldmann), Nora K. Speicher, Prabhav Kalaghatgi, and Tomas Bastys. Thank you.

Dedicated to my father, Klaus-Peter († 2012)

Abstract

Epigenetics is the field of biology that investigates heritable factors regulating gene expression without being directly encoded in the genome of an organism. The human genome is densely packed inside a cell's nucleus in the form of chromatin. Certain constituents of chromatin play a vital role as epigenetic factors in the dynamic regulation of gene expression. Epigenetic changes on the chromatin level are thus an integral part of the mechanisms governing the development of the functionally diverse cell types in multicellular species such as human. Studying these mechanisms is not only important to understand the biology of healthy cells, but also necessary to comprehend the epigenetic component in the formation of many complex diseases.

Modern wet lab technology enables scientists to probe the epigenome with high throughput and in extensive detail. The fast generation of epigenetic datasets burdens computational researchers with the challenge of rapidly performing elaborate analyses without compromising on the scientific reproducibility of the reported findings. To facilitate reproducible computational research in epigenomics, this thesis proposes a task-oriented metadata model, relying on web technology and supported by database engineering, that aims at consistent and human-readable documentation of standardized computational workflows. The suggested approach features, e.g., computational validation of metadata records, automatic error detection, and progress monitoring of multi-step analyses, and was successfully field-tested as part of a large epigenome research consortium. This work leaves aside theoretical considerations, and intentionally emphasizes the realistic need of providing scientists with tools that assist them in performing reproducible research.

Irrespective of the technological progress, the dynamic and cell-type specific nature of the epigenome commonly requires restricting the number of analyzed samples due to resource limitations. The second project of this thesis introduces the software tool SCIDDO, which has been developed for the differential chromatin analysis of cellular samples with potentially limited availability. By combining statistics, algorithmics, and best practices for robust software development, SCIDDO can quickly identify biologically meaningful regions of differential chromatin marking between cell types. We demonstrate SCIDDO's usefulness in an exemplary study in which we identify regions that establish a link between chromatin and gene expression changes. SCIDDO's quantitative approach to differential chromatin analysis is user-customizable, providing the necessary flexibility to adapt SCIDDO to specific research tasks.

Given the functional diversity of cell types and the dynamics of the epigenome in response to environmental changes, it is hardly realistic to map the complete epigenome even for a single organism like human or mouse. For non-model organisms, e.g., cow, pig, or dog, epigenome data is particularly scarce. The third project of this thesis investigates to what extent bioinformatics methods can compensate for the comparatively little effort that is invested in charting the epigenome of non-model species. This study implements a large integrative analysis pipeline, including state-of-the-art machine learning, to transfer chromatin data for predictive modeling between 13 species. The evidence presented here indicates that a partial regulatory epigenetic signal is stably retained even over millions of years of evolutionary distance between the considered species. This finding suggests complementary and cost-effective ways for bioinformatics to contribute to comparative epigenome analysis across species boundaries.

Kurzfassung

Epigenetik ist das Teilgebiet der Biologie, welches vererbbare Faktoren untersucht, die die Genexpression regulieren, ohne dabei direkt im Genom eines Organismus kodiert zu sein. Das menschliche Genom liegt dicht gepackt im Zellkern in der Form von Chromatin vor. Bestimmte Bestandteile des Chromatin spielen als epigenetische Faktoren eine zentrale Rolle bei der dynamischen Regulation von Genexpression. Epigenetische Veränderungen auf Chromatinebene sind daher ein integraler Teil jener Mechanismen, die die Entwicklung von funktionell diversen Zelltypen in multizellulären Spezies wie Mensch maßgeblich steuern. Diese Mechanismen zu untersuchen ist nicht nur wichtig, um die Biologie von gesunden Zellen zu erklären, sondern auch, um den epigenetischen Anteil an der Entstehung von vielen komplexen Krankheiten zu verstehen.

Moderne Labortechnologien erlauben es Wissenschaftlern, Epigenome mit hohem Durchsatz und sehr detailliert zu erforschen. Ein schneller Aufbau von epigenetischen Datensätzen stellt die computerbasierte Forschung vor die Herausforderung, schnell aufwendige Analysen durchzuführen, ohne dabei Kompromisse bei der wissenschaftlichen Reproduzierbarkeit der gelieferten Ergebnisse einzugehen. Um die computerbasierte reproduzierbare Forschung im Bereich der Epigenomik zu vereinfachen, schlägt diese Dissertation ein aufgabenorientiertes Metadaten-Modell vor, welches, aufbauend auf Internet- und Datenbanktechnologie, auf eine konsistente und gleichzeitig menschenlesbare Dokumentation für standardisierte computerbasierte Arbeitsabläufe abzielt. Das vorgeschlagene Modell ermöglicht unter anderem eine computergestützte Validierung von Metadaten, automatische Fehlererkennung, sowie Fortschrittskontrollen bei mehrstufigen Analysen, und wurde unter realen Bedingungen in einem epigenetischen Forschungskonsortium erfolgreich getestet. Die beschriebene Arbeit präsentiert keine theoretischen Betrachtungen, sondern setzt den Schwerpunkt auf die realistische Notwendigkeit, Forscher mit Werkzeugen auszustatten, die ihnen bei der Durchführung von reproduzierbarer Arbeit helfen.

Unabhängig vom technologischen Fortschritt, erfordert die zellspezifische und dynamische Natur des Epigenoms häufig eine Beschränkung bei der Anzahl an zu untersuchenden Proben, um Ressourcenvorgaben einzuhalten. Das zweite Projekt dieser Arbeit stellt die Software SCIDDO vor, welche für die differenzielle Analyse von Chromatindaten auch bei geringer Verfügbarkeit von Zellproben entwickelt wurde. Durch die Kombination von Statistik, Algorithmik, und bewährten Methoden zur robusten Software-Entwicklung, erlaubt es SCIDDO, schnell biologisch sinnvolle Regionen zu identifizieren, die ein differenzielles Chromatinprofil zwischen Zelltypen aufzeigen. Wir demonstrieren SCIDDOs Nutzwert in einer beispielhaften Studie, z.B. durch die Identifikation von Regionen, die eine Verbindung von Änderungen auf Chromatinebene und Genexpression herstellen. SCIDDOs quantitativer Ansatz bei der differenziellen Analyse von Chromatindaten erlaubt eine nutzer- und aufgabenspezifische Anpassung, was Flexibilität bei der Bearbeitung anderer Fragestellungen ermöglicht.

Bedingt durch die funktionelle Vielfalt an Zelltypen und die Dynamik des Epigenoms resultierend aus Umgebungsveränderungen, ist es kaum realistisch, das komplette Epigenom von auch nur einer einzigen Spezies wie Mensch zu erfassen. Insbesondere für nicht-Modellorganismen wie Kuh, Schwein, oder Hund sind sehr wenig Epigenomdaten verfügbar. Das dritte Projekt dieser Dissertation untersucht, inwieweit bioinformatische Methoden dazu verwendet werden könnten, den vergleichsweise geringen Aufwand, welcher betrieben wird um das Epigenom von nicht-Modellspezies zu erforschen, zu kompensieren. Diese Studie realisiert eine große, integrative Computeranalyse, welche basierend auf Methoden des maschinellen Lernens und auf Transfer von Chromatindaten Modelle zur Genexpressionsvorhersage über Speziesgrenzen hinweg etabliert. Die gewonnenen Erkenntnisse lassen vermuten, dass ein Teil des regulatorischen epigenetischen Signals auch über Millionen von Jahren an evolutionärer Distanz zwischen den 13 betrachteten Spezies stabil erhalten bleibt. Diese Arbeit zeigt dadurch ergänzende und kosteneffektive Möglichkeiten auf, wie Bioinformatik einen Beitrag zur vergleichenden Epigenomanalyse über Speziesgrenzen hinweg leisten könnte.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Lead-in This chapter is written as a gentle introduction into the central topics of this thesis. The depth of this chapter is thus intentionally limited. The remaining chapters present more detailed background information, which permits to focus here on selected examples. Readers already familiar with scientific work in biological research can probably skip the first two sections and proceed directly to the thesis outline without missing vital pieces of information.

1.1 Comparative Observation and Reproducibility by Example

Arguably one of the biggest leaps forward in the field of biology was initiated by Charles Darwin's "On the Origin of Species" [54]. In his seminal work, Darwin argues that biological complexity develops gradually over time, and species do not come into existence by independent creation, but by successively acquiring new traits or losing old ones relative to their ancestral generation(s). This process, free of any planning or foresight, is driven by natural selection of advantageous — as opposed to deleterious — variants of biological characteristics, and results in species that are well-adapted to their ecological niche. Darwin's elegant model explaining observed biological variation as an evolutionary process driven by natural selection also builds on two core concepts of scientific work, both of which are central to the work presented in this thesis: comparative observation and scientific reproducibility.

Comparative observation, i.e., examining similarities and differences between different subjects under study, represents a common approach toward viewing the world with a scientific mindset. Although Darwin was of course not the first naturalist to keenly observe life around him, he made the important next step of formulating a complete and biologically plausible hypothesis why certain similarities exist in visibly distinct animal species. A common textbook example for such an observed similarity is the forelimb of many vertebrate species that all share the same, so-called tetrapod, bone structure (Figure 1.1) [253]. According to Darwin's theory of evolution, these different forms of the same basic layout developed from a common ancestor by natural selection of those naturally occurring variants that provided an advantage in the struggle for reproduction. The tetrapod bone structure in the forelimb is called a homologous trait, i.e., a trait that has developed from a common ancestor in the respective vertebrate species.
Roughly 100 years after Darwin shared his theory of evolution with the public, technological

advances enabled scientists to explore the molecular aspects of biology and to reason about the presumed universality of the genetic code [130]. A universal[1] genetic code implies, under Darwin's theory of evolution, that there should be a last universal common ancestor (LUCA) from which all living organisms on this planet originated. From that, it immediately follows that genome sequences can be homologous in nature in the same way as, e.g., the bone structure in vertebrate forelimbs. This realization gave rise to the field of comparative genomics, i.e., the field that studies similarities and differences of genomic features across organisms. The experimental methods in comparative genomics have probably nothing in common with those of Darwin's era, yet the concept of learning through comparative studies still applies and is the approach taken in Chapters 4 and 5 of this thesis.

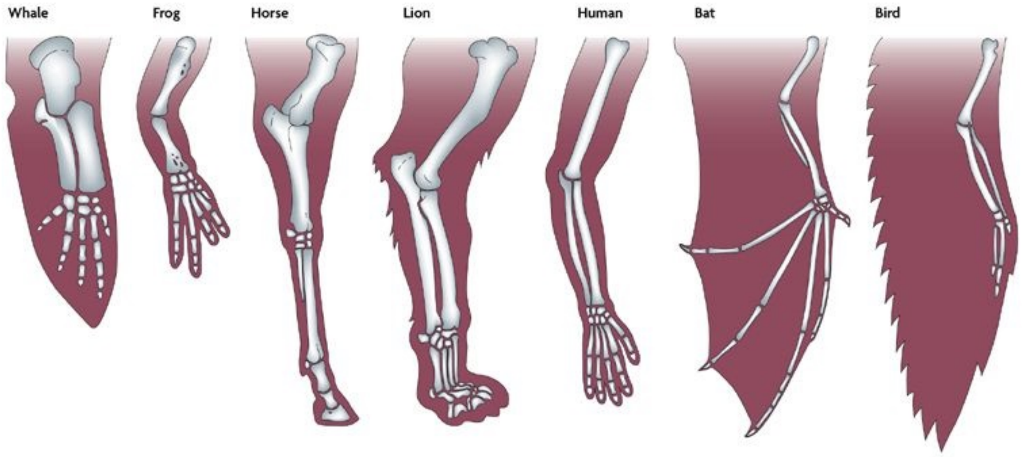


Figure 1.1: The tetrapod bone structure in vertebrates: similar bone structure in the forelimbs of several tetrapod vertebrate species exemplifies the evolutionary principle of gradual modification of a trait that emerged in a common ancestor. Functional necessity cannot explain the observed similarities given the different purposes of the forelimbs, e.g., swimming and flying. The tetrapod bone structure is a homologous trait in the respective vertebrate species (license #4432540824975, see Table E.1).

The second concept is essential to the "scientific survival" of Darwin's theory and is implicitly given in the following statement:

> "If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down. But I can find out no such case."
>
> (cf. Chapter 6 "Difficulties On Theory - Organs of extreme perfection")

What Darwin is referring to here is the problem that he cannot prove his theory of evolution to be true, but as long as no contradicting evidence is presented, one can think of the theory as "not false". In more general terms, any scientific hypothesis must be independently testable or withstand a re-examination by third parties. If such independent testing achieves the same (observational)

[1] A recent publication describes an exception to this universality of the genetic code [168]. For the sake of brevity, this will not be discussed here.

conclusions as originally reported, the hypothesis is less likely to be a result of, e.g., chance observations. This line of thinking is central to the concept of scientific reproducibility that will be discussed in more detail in Chapter 3. The fact that, to this day, Darwin's theory of evolution is considered valid means that not only could Darwin himself not find a counterexample, but also that nobody else could. In other words, in the approximately 160 years since Darwin published "On the Origin of Species", no conclusive evidence could be collected that would allow to refute Darwin's theory. This long-lasting accumulation of evidence in favor of the theory could explain why Darwin's theory of evolution is commonly (mis-) represented as an absolute truth in mainstream media; from an epistemological point of view, this is not correct, but it seems plausible to assume that any scientist would like to see their hypotheses advance to that level of perceived certainty. In a broader sense, and as will be argued in Chapter 3, progress in a scientific field requires a reproducible basis of knowledge — first principles perceived as "known to be true" — as starting point for future studies.

1.2 High-Throughput Biology and Epigenetics

Jumping forward from the early days of comparative genomics in the second half of the 20th century to the turn of the millenium, a breakthrough was achieved through the sequencing of the human genome, with a first draft sequence publicly released almost 20 years ago [50, 136, 247]. This achievement had many implications, of which two shall be mentioned in the context of this thesis. First, knowing the sequence of the human genome enables scientists to identify genomic sequence homology relative to other species, thereby illuminating the traces of evolution in our genome:

> "More generally, comparative genomics allows biologists to peruse evolution's laboratory notebook — to identify conserved functional features and recognize new innovations in specific lineages. [...] Plans are also under consideration for sequencing additional primates and other organisms that will help define key developments along the vertebrate and non-vertebrate lineages. To realize the full promise of comparative genomics, however, it needs to become simple and inexpensive to sequence the genome of any organism."
> (Lander et al. [136])

The latter point about economic limitations is no longer a major issue for large-scale sequencing projects due to the rise of comparatively cheap and scalable technologies collectively referred to as next–generation sequencing (NGS) or high-throughput sequencing (HTS) technologies [163]. This technological advance also fuels large scientific projects with the objective of sequencing thousands of different animal species [128]. Hence, scientists in (comparative) genomics are now at a point where they can create large sequence databases for a multitude of different species. This recent development has enabled cross-species studies such as the one presented in Chapter 5 of this thesis.

The second implication of the completed human genome sequence is much more far-reaching, but it is also much harder to assess its potential impact on biological and biomedical research:

> “In principle, the string of genetic bits *(note: the sequenced genome)* holds long-sought secrets of human development, physiology and medicine. [...] Fulfilling the true promise of the Human Genome Project will be the work of tens of thousands of scientists around the world, in both academia and industry. [...] The scientific work will have profound long-term consequences for medicine, leading to the elucidation of the underlying molecular mechanisms of disease and thereby facilitating the design in many cases of rational diagnostics and therapeutics targeted at those mechanisms. [...] We must set realistic expectations that the most important benefits will not be reaped overnight.”
> (Lander et al. [136])

Though it would be an absurd question to ask how long it will take until the above mentioned benefits, especially in the area of medicine, are realized, it seems safe to say that “20 years” would not have been the right answer. To illustrate that point, one might consider the exemplary case of cancer. Substantial progress in diagnosis and treatment has been made for certain types of cancer also as a consequence of advances in modern genomics, yet cancer is far from being generally understood and “under control”, i.e., despite increasing survival rates, cancer is still among the leading causes of death worldwide [93, 94][2]. One important aspect to the question why knowing the human genome is not enough to, e.g., cure diseases such as cancer, is within the reach of basic biology: the genome exists essentially in form of identical copies in almost every cell of the body, since, at the beginning of human life, there is only a single fertilized egg. All other cells in the human body ultimately derive their copy of the genome from this ancestral cell. It seems compelling that there has to be a process that controls how hundreds of different cell types in the human body can be created from a single starting point. The field investigating these processes is called epigenetics. The founding father of epigenetics, Conrad H. Waddington, described it as “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (quoted after Goldberg et al. [89]). A more contemporary definition of epigenetics is given by Goldberg et al. as follows: “[...] epigenetics may be defined as the study of any potentially stable and, ideally, heritable change in gene expression or cellular phenotype that occurs without changes in Watson-Crick base-pairing of DNA”[3]. In other words, epigeneticists study how the information in our genome is turned into functionally different cell types, and what deviations from normal regulation constitute a disease state, e.g., as observed in many cancers [74, 114].

Although Conrad H. Waddington coined the term epigenetics decades ago, and its relevance for the development of the cellular phenotype is no recent discovery, the fact that so many different and dynamic factors contribute to the epigenome of a cell prohibited larger epigenome

[2] cancer.gov/about-cancer/understanding/statistics

[3] This thesis does not elaborate on different standpoints regarding a “precise” definition of epigenetics. Section 2.1.1.5 in Chapter 2 is the only part of this thesis where a definition of epigenetics emphasizing the heritability is relevant. Some literature references discussing issues on the definition and use of the term epigenetics are thus included in Section 2.1.1.5.

studies until the advent of cheap and scalable NGS technologies. In that sense, (high-throughput) epigenomics is still an emerging field that faces many fundamental challenges, e.g., establishing experimental protocols, defining data standards, or developing common vocabulary. Such tasks are rather technical in nature, but they are needed to provide the proper footing for future studies. Despite technicalities that need to be addressed, research in epigenomics holds great potential for advancing both basic research and more applied fields such as biomedicine [7, 23, 27, 89, 179, 198].

In the sections above, the anchoring points of this thesis have been introduced with some historical pointers: scientific reproducibility and comparative analyses in the field of (epi-) genomics. The following section now provides a detailed outline of the remainder of this thesis, and summarizes central points and research directions of each chapter.

1.3 Research Directions and Outline of this Thesis

In Chapter 2, the biological and computational background is introduced. This background information focuses on the part of epigenomics and bioinformatic data processing that is required as basic knowledge for the three methodological chapters 3, 4, and 5. Each of these chapters in turn includes a section that details chapter-specific data and method information.

Chapter 3 deals with the question of how scientific reproducibility can be practically integrated into the research routines of large collaborative projects, which are a prevalent organizational form in modern biological research [56, 106, 136, 202, 233]. Issues with scientific reproducibility are an area of active debate [4, 73, 171, 172, 192], and large research initiatives, with their substantial economic and scientific means, hold the role of key players in terms of setting experimental standards. Hence, research initiatives also carry part of the responsibility to consolidate work in emerging fields such as epigenomics. Any work supporting efforts of conducting research in a reproducible way is thus an important contribution to the overall level of quality in a research project.

Specifically, the approaches presented in Chapter 3 were prototypically developed and field-tested as part of the German epigenome programme (Deutsches Epigenom Programm [48] (DEEP)). The work in Chapter 3 starts with a brief introduction on scientific reproducibility and a summary of several current approaches of how scientific work is conducted in a reproducible way. Based on this background, the strategies and the software tools combining web technology and databases that were developed in DEEP are described. The evaluation part of Chapter 3 highlights the advantages of the implemented approach and discusses several observed shortcomings. Chapter 3 concludes with an empirically motivated attempt to derive some modest principles from the lessons learned in DEEP that could be implemented in the future to gradually improve reproducibility rates in computational research.

The next two chapters rely conceptually on the foundations of Chapter 3. Assuming that data generation and basic data processing are established in reproducible ways, computational research projects can turn to more biology-oriented questions.

The central question of Chapter 4 is if it is possible to identify genomic regions that show a distinct epigenetic profile between different cell types within one species. The type of data analyzed in Chapter 4 is an abstract representation of an epigenetic signal that has been shown to provide a rich view on regulatory processes in the cell [29, 72, 119, 205]. In the affirmative case, the identified regions could thus be specific for the respective cellular phenotype if the regions could additionally be shown to be biologically plausible and meaningful. The software SCIDDO has been developed to tackle the question of Chapter 4 by combining statistical evaluation procedures with fast algorithms in a lightweight command line tool. In the case study presented in Chapter 4, SCIDDO could identify such regions in a medium-sized real-world dataset within minutes. The further biological characterization of the identified regions provides evidence that the epigenetic profile in these regions conforms to the current understanding of epigenetic regulation in cellular processes, e.g., changes on the chromatin level can be linked to changes in gene expression patterns. SCIDDO is the first software that implements a score-based approach for the identification of the genomic regions of interest. This score-based approach provides a flexible and user-customizable way of emphasizing different aspects of epigenetic regulation. Biological plausibility of the identified regions, processing speed and flexibility of SCIDDO let us conclude that SCIDDO is a valid contribution to the bioinformatics toolbox in computational epigenomics research.

Chapter 5 presents an exploratory study that extends epigenome-based comparative analyses to non-model vertebrate species. The work in Chapter 5 explores if it is possible to transfer knowledge acquired in epigenetically well-characterized animals such as human or mouse to other species that are not canonical model organisms, and hence have a largely uncharted epigenome. As opposed to the analytical setting in Chapter 4, which builds on abundant epigenome datasets in human and thus can rely on an abstract representation of biological information, the idea of Chapter 5 is motivated by the problem that epigenomics is a resource-intensive field of biology. Consequently, only comparatively few and hardly comprehensive epigenome resources exist for animal species that are not established model organisms for epigenomics research. However, factors such as the economic value of species like cow and pig, or the potential medical value of vertebrate species having an exceptionally low incidence rate of diseases such as cancer [91, 120, 246] make cross-species studies involving non-model organisms a worthwhile endeavor. The entire approach in Chapter 5 is performed *in silico* and sidesteps any wet lab resource limitations. The results indicate that it is possible to transfer epigenetic data across species boundaries and still retain sufficient information to perform predictive modeling on the transferred data using common machine learning tools. In Chapter 5, the machine learning-based prediction of cellular characteristics across evolutionary distances of more than 300 million years is demonstrated to be possible with acceptable accuracy. Hence, this study transcends the idea of comparative observation by testing if observations can, at least partially, be replaced with computational estimates. The principle validity of this approach rests on the already described insight that shared ancestry of organisms should result in some discoverable similarities; in the context of Chapter 5, these presumed similarities are epigenetic in nature.

This thesis concludes with Chapter 6, which summarizes and interrelates the three methodological chapters outlined above and discusses potential future work. Chapter 6 does not, however, reiterate the concluding remarks of Chapter 3 (Section 3.6), but highlights some aspects of ongoing collaborative research efforts in the field of epigenomics that fit the context of Chapter 3.

CHAPTER 2

Background

Lead-in This chapter provides a general overview of epigenetics with a focus on histone proteins and their role in regulation of cellular processes. This chapter ends with a description of how the biological histone signal is translated into computer-readable information, and then processed in bioinformatics pipelines as implemented by many national and international research consortia. This standardized way of processing histone data is relevant in the context of Chapter 3. In Chapters 4 and 5, histone data are used as mere input to bioinformatics analyses, and thus the discussion of potential sources of noise presented in Section 2.3.1 is of implicit importance in these chapters.

2.1 (Epi-) Genetics and Chromatin

The biological concepts discussed in this chapter require some basic knowledge in genetics for easy comprehension. The next paragraph introduces these definitions in a non-exhaustive manner focusing on the aspects relevant for this thesis:

As highlighted in Chapter 1, one of the central questions that is studied in epigenetics is how fine-tuned epigenetic regulation contributes to the development of the functionally diverse cell types in multicellular species such as human. “Epigenetic regulation” refers to the processes that modulate the transcriptional or regulatory state in confined regions of the genome. This modulation can be thought of as either absolute, e.g., switching from an inactivate to an active state, or as gradual, e.g., increasing or decreasing local activity levels. Transcription usually refers to the transcription of deoxyribonucleic acid (DNA) into ribonucleic acid (RNA) in a genomic region that represents the locus of a protein-coding gene. Since a precise definition of what a gene is, seems to be elusive [85], a “working definition” of a protein-coding gene is sufficient here: a gene is a heritable DNA segment that codes for at least one protein and that is associated with “regulatory regions”. These regulatory regions are important, e.g., for gene activation and initiation of transcription (“gene promoter”), or for the regulation of the gene’s activity level (“gene-associated enhancer”). The activity of a gene is usually quantified as the gene expression level, where expression refers here to the process of transcribing the nucleotide sequence of a gene into messenger RNA.

Following the main topics of this thesis, the description of epigenetic factors and their role in regulation given below focuses on the normal or healthy cellular condition; detailed descriptions of aberrant patterns of epigenetic regulation, e.g., related to disease phenotypes, are omitted. RNA-mediated ways of epigenetic transcriptional control and RNA constituents of chromatin are not covered as they are not relevant for the following chapters.

The name "chromatin" was coined by Walther Flemming in the late 19th century when he realized that the scaffold inside a cell's nucleus could easily be stained (*chroma* is Greek for color; chromatin can thus be translated to "stainable material") [181]. The related term "chromosome" was introduced around 10 years later by Heinrich Wilhelm Waldeyer, and translates to "stainable body", i.e., a constituent component of chromatin [181]. In more recent literature, chromatin was described as consisting of DNA, RNA and proteins, and the term refers collectively to all chromosomes [167]. The realization that "[i]n such chromatin, only a portion of the genetic material is available for transcription by RNA polymerase, and the genes thus accessible are the same ones that are accessible and transcribed in life" [30] is immediately suggestive of one of the major characteristics of chromatin: while chromatin can be considered a "packaging scaffold" to fit the approximately two meters of DNA inside the cell's nucleus, chromatin also has to be locally flexible enough to expose certain parts of the genome, e.g., to enable proper gene transcription. These two states of chromatin — open and accessible or closed and inaccessible — are called euchromatin (open) and heterochromatin (closed). The flexibility to change from heterochromatin to euchromatin in confined regions of the genome is possible due to the hierarchical structure of chromatin. At the basic level, approximately 150 base pairs (bp) of DNA are wrapped around histone proteins and this complex of DNA and histone proteins is called a nucleosome. The serial arrangement of nucleosomes along the genome is commonly referred to as the "beads on a string" model (see Figure 2.1, second row), and more and more stages of twisting and coiling eventually result in the necessary level of compaction to fit the DNA into the nucleus (Figure 2.1, top to bottom). While it may appear that the structure of chromatin is highly organized, the recent characterization of so-called nucleosome clutches by Ricci et al. suggests otherwise [204]. Ricci et al. defined nucleosome clutches as groups of nucleosomes that show variation in size and density in a cell-type specific manner. This high-level view on chromatin organization hints at an additional layer of detectable cell-type specificity that, for the sake of brevity, will not be discussed here. Instead, the following section focuses on the histone proteins.

2.1.1 Histone proteins and nucleosomes

Histone proteins are divided into two super families and five subordinate families: the linker histone family H1, and the four core histone families H2A, H2B, H3 and H4. Two members of each core family build the octamer core of a nucleosome (Figure 2.2) [151], whereas the H1 histones attach to the linker DNA that connects adjacent nucleosomes (cf. Figure 2.1, second row). There is a considerable number of histone variants in each family, and the occurrence of many of these variants has been linked to, e.g., specific phases of the cell cycle, to certain eukaryotic lineages, or to disease phenotypes [159, 234, 235]. In the context of this work, histone variants are not relevant

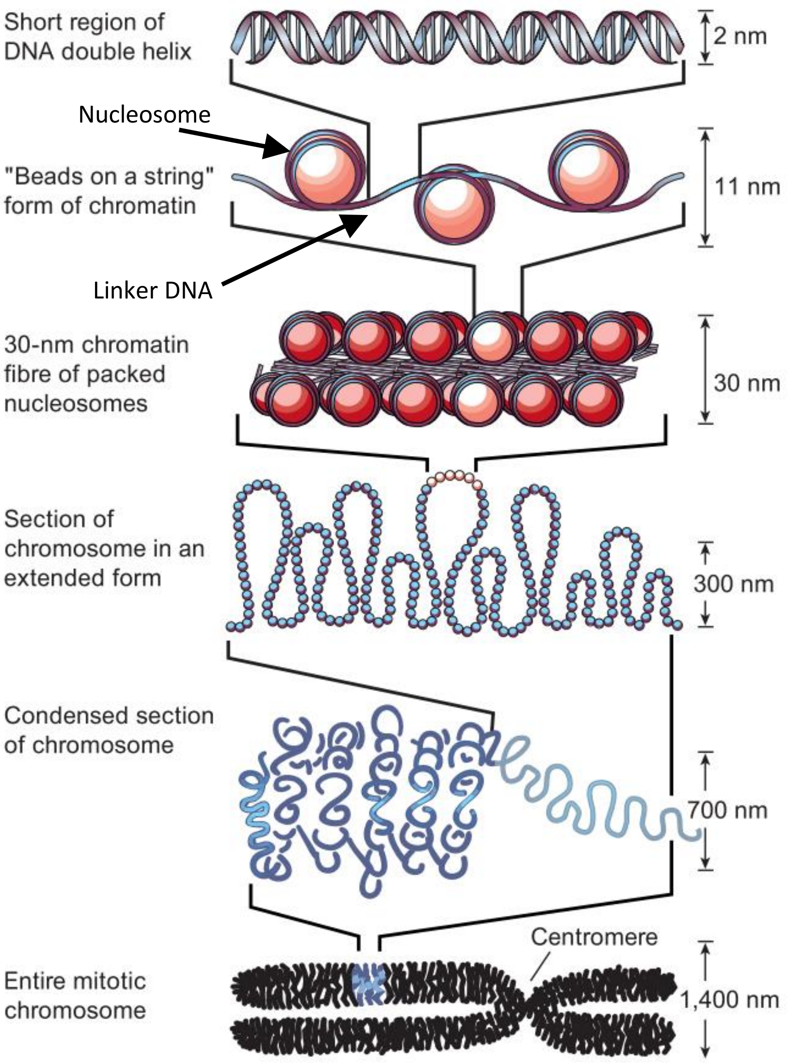

Short region of DNA double helix
2 nm
Nucleosome
"Beads on a string" form of chromatin
11 nm
Linker DNA
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

Figure 2.1: Packaging of DNA in chromatin: chromatin organization inside the cell's nucleus has a hierarchical order, starting with individual nucleosomes separated by short stretches of linker DNA (second row); the linker histone H1 bound to the linker DNA is not shown. Precise chromatin organization beyond the 11 nanometer (nm) fiber ("beads on a string" model) is still an area of active research [204] (license #4407631152362, see Table E.1; labels "Nucleosome" and "Linker DNA" and corresponding arrows were manually edited into the figure).

and it suffices to note that histone variants represent a source of epigenetic variation in addition to those discussed in this chapter (in particular Section 2.1.1.4).

The four core histones are small proteins (~100 to ~135 amino acids) that have three α-helical domains and an unstructured N-terminal domain, referred to as histone tail [151]. The individual amino acids of a histone protein are identified by their position relative to the N-terminus, e.g., the first lysine of histone 3 is located at position 4, and hence identified by the shorthand notation "H3K4". This shorthand notation is used throughout the remainder of this thesis. When a H3-H4 tetramer and two H2A-H2B dimers are assembled to form the octameric core of a nucleosome, the histone tails protrude from the center of the nucleosome and are amenable to chemical modifications by other enzymes (Figure 2.2); the role of these post-translational modifications is discussed in Section 2.1.1.4. As pointed out above, the spatial organization of nucleosomes is not necessarily rigid, implying that nucleosome formation, at least at certain genomic locations, has to be a regulated process with some discernible specificity.

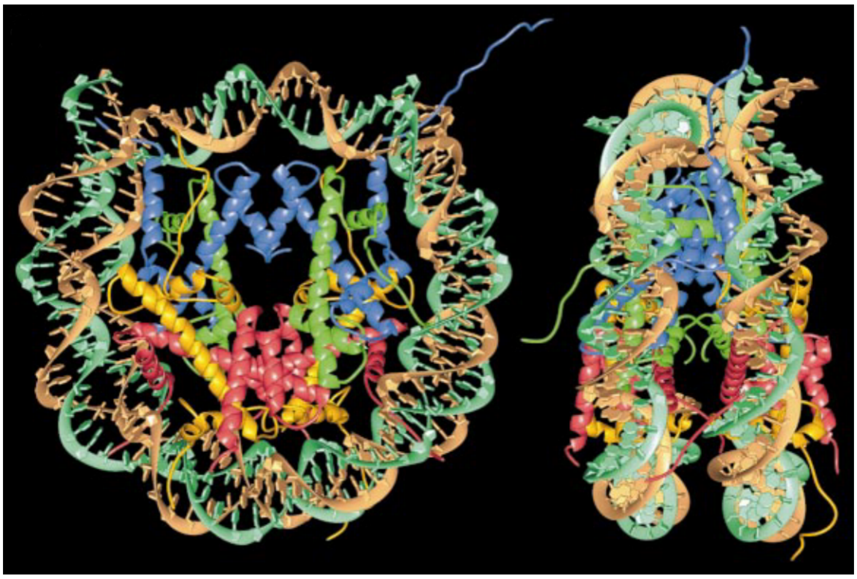


Figure 2.2: Ribbon traces of the nucleosome core particle showing 146 bp of the DNA backbone and the four core histone proteins (blue: H3; green: H4; yellow: H2A; red: H2B). The histone tails visibly protrude from the core and are thus accessible for other enzymes that bind or modify individual residues (see Section 2.1.1.4 below; license #4407621492444, see Table E.1).

2.1.1.1 Histone deposition and nucleosome formation

Before some principles of nucleosome formation are introduced, it is helpful to elucidate two related terms that are often used interchangeably and lack a community-wide accepted standard definition [223]: nucleosome positioning and nucleosome occupancy. Nucleosome positioning refers to the position of a nucleosome along the genome sequence. A “highly positioned” or “strongly positioned” nucleosome could in principle be identified by its “sequence coordinates”, i.e., the combination of nucleotide position(s) and chromosome number. Instead, relative nucleosome coordinates are commonly used. For example, the strongly positioned nucleosome immediately downstream of the transcription start site (TSS) is conventionally identified as the “+1” nucleosome. The number of strongly positioned nucleosomes has been reported to be around 10%, with many more nucleosomes showing at least a tendency toward consistent genomic positions [82].

Nucleosome occupancy, on the other hand, refers to the presence or absence of a nucleosome in specific sequence contexts. For example, when we examine a stretch of DNA that has a strongly positioned nucleosome and a high occupancy, we would expect to find a nucleosome at exactly this location every time we examine the same stretch of DNA again. However, if the occupancy were low, we can only say that, if we find a nucleosome, it will be located at exactly the same position. This consideration is relevant for the interpretation of the high-throughput assays introduced in Section 2.2, where the biological material for an experiment is not just a single cell, but a large quantity of functionally homogeneous cells.

The process of nucleosome formation requires the stepwise assembly and deposition of the H3-H4 tetramer or of H2A-H2B dimers onto the DNA, resulting in pre- or immature nucleosomes. These nucleosome building blocks are delivered by histone chaperones such as CAF1, which are broadly conserved throughout the eukaryotic lineage, but the process of initial assembly is not yet understood in all details [158]. Another class of proteins, so-called chromatin remodelers, are responsible for finalizing histone core assembly and for creating regularly spaced, mature nucleosomes [44]. Nucleosome assembly is a process that directly follows DNA replication (see also Section 2.1.1.5) and, in this context, it may not seem obvious why nucleosome positioning would have to be regulated, since the information is essentially just copied. However, given the observation of strongly and weakly positioned nucleosomes and of nucleosome mobility (Section 2.1.1.2), the existence of factors determining nucleosome positions seems compelling. It is generally assumed that there is some sequence preference, motivated by the fact that the DNA bending around the histone core favors periodic occurrences of certain dinucleotides for their physical properties [151]. Moreover, the characterization of sequences — most notably the Widom 601 sequence [148] — exhibiting a strong nucleosome positioning effect (*in vitro*) similarly suggested a sequence-dependent positioning process. However, it seems implausible that predicting (all) nucleosome positions from sequence alone would be reasonable for functionally diverse cell types in multicellular organisms, as the predictions would necessarily be tissue-independent and could thus not capture the observed flexibility in nucleosome positioning/occupancy. Nevertheless, some studies suggest that sequence-based (prediction) models could be helpful for better characterizing the role of sequence effects relative to other factors (e.g., [86, 245]). More comprehensive but abstract models of nucleosome positioning and occupancy also consider factors like the local regulatory environment, higher order chromatin structure, and the effect of potentially competitive binding by transcription factors (TFs) [223]. Under these conditions, nucleosomes may be placed or moved into sequence contexts with unfavorable physical properties, which is only possible because other factors compensate for this unfavorable sequence context. Despite the existence of strongly positioned nucleosomes, it seems reasonable to view nucleosomes as rather dynamic entities, in particular during important cellular processes such as gene transcription.

2.1.1.2 Nucleosome mobility

The passage of the transcriptional machinery along the gene body is occluded by nucleosomes that are particularly well-positioned in exons [9]. There are at least two obvious solutions to this problem: the nucleosomes have to be either disassembled (sometimes referred to as disruption or eviction) and reassembled after the transcriptional machinery has passed, or the tight contact between the DNA and the histone core has to be relaxed to make the DNA accessible for the transcriptional machinery. The observation that the canonical H3 histone variant is replaced by the variant H3.3 in transcribed genes provided supporting evidence for the first solution [260]. Since H3.3 is not produced in the synthesis phase of the cell cycle when DNA is replicated, and thus could not have been incorporated into the histone core of the nucleosome as described in Section 2.1.1.1, it is assumed that some nucleosomes are indeed completely disrupted and reassembled during transcription.

The second way of making the DNA accessible involves only partial disassembly of nucleosomes. The FACT (for "facilitates chromatin transcription") complex has been shown to displace H2A-H2B dimers of the histone core, which destabilizes the nucleosome and makes the DNA accessible enough for the transcriptional machinery to proceed [16, 17]. Besides this partially disruptive process, chromatin remodeling complexes can also "slide" nucleosomes along the DNA following the so-called "wave-ratchet-wave" model [44, 95, 209]. In this model, a chromatin remodeling complex binds to the nucleosomal DNA at a specific location, and pulls in DNA from one side of the nucleosome, locally breaking histone-DNA contacts and forming a small bulge of DNA. The resulting tension in the DNA is released by propagating the DNA bulge toward the other end of the nucleosome in a wave-like fashion at a step size of around 1–2 bp.

Whereas the preceding paragraph deals with mechanisms that restore the local chromatin environment after a temporary alteration, the next section exemplifies cellular processes that remove nucleosomes permanently.

2.1.1.3 Nucleosome eviction

In scenarios in which the local chromatin environment is in closed conformation and needs to be accessible, e.g., when a gene is activated and the promoter region has to be opened to enable transcriptional initiation, a complete removal of specific nucleosomes has been observed. In this case, additional signals are required to trigger the eviction process, e.g., initiated by TF binding [140, 260]. To give an example, one of those TFs is the acetyl transferase EP300 (p300), which acetylates lysine 14 of histone H3 (H3K14), leading to subsequent destabilization and eviction of the modified nucleosome by the histone chaperone Nap1 [150]. This process opens some space in the promoter region and enables transcriptional initiation. A similar destabilizing effect of histone acetylation has been reported by Chatterjee et al. [40], who observed that nucleosomes with an acetylated histone H3 (lysines at position 115 and 122) have a higher predisposition of being disassembled by chromatin remodeling complexes. Another mechanism specifically deposits nucleosomes carrying the histone variant H2A.Z in promoters of inactive genes. Whereas these nucleosomes are rather resistant against chromatin remodeling, they are removed quickly upon transcriptional activation due to their lower stability compared to nucleosomes with canonical H2A [140].

It should be pointed out that nucleosome eviction in the context of transcriptional activation does not lead to promoters that are entirely devoid of nucleosomes. As a consequence, gene promoters — and other regulatory elements in the genome — are amenable to basic classification schemes based on histone post-translational modifications that commonly occur in the respective regulatory contexts.

2.1.1.4 Histone post-translational modifications

Histone proteins can be post-translationally modified by chromatin modifying enzymes (not to be confused with chromatin remodeling enzymes). The two chemical modifications, histone marks for short, that are important in the context of this work are acetylation and methylation of lysine residues; other modifications will not be discussed here [132]. The enzymes facilitating post-translational modification of histone residues, i.e., lysine methyltransferase (KMT), lysine demethylase (KDM), lysine acetyltransferase (KAT) and histone deacetylase (HDAC) [6], plus those enzymes that bind to modified residues are often collectively referred to has "readers" and "writers/erasers" of histone marks, presumably motivated by the notion of a "histone code" [110]. The histone code hypothesis builds on the following line of thought: histone marks, or rather combinations of them, constitute recognition sites for other chromatin-associated proteins with potentially synergistic or antagonistic effects. These other chromatin-associated proteins in turn regulate changes in genomic activity levels, e.g., switching from silent to transcriptionally active chromatin states. Hence, the histone code hypothesis postulates that the combinatorial space of all histone marks extends the information encoded by the sequence of nucleotides in the genome. Large-scale studies of dozens of histone acetylation and methylation marks identified a core set of 17 histone marks that seem to be more or less ubiquitously present in promoter regions of transcribed genes, suggesting that there could be a histone signal strictly indicating active transcription [255]. However, in the same study, approximately 75% of all detected combinations of histone marks were only present at one gene each. While these numbers cannot be taken at face value, as Wang et al. also noted, due to noise in the data, they seem to suggest that, if there is a histone code, it would probably be quite challenging to clearly identify distinct "code words". Taking this unresolved state of the histone code into account, and because "reading" and "writing" histone marks provokes associations of language and interpretation, histone-modifying enzymes (HMEs) will not be referred to as "readers" and "writers/erasers" in the remainder of this thesis. Instead, those enzymes that set or remove a chemical modification will be referred to as HMEs, and those enzymes that have a recognition domain for specific marks will be referred to as histone-binding domain enzymes (HBDEs).

Current literature lists more than one hundred different histone marks and reports on their functional role implicating them in a large variety of regulatory processes [123, 132]. To limit the complexity, it is common to focus on a set of "core marks" that are broadly representative of many important chromatin activity levels ranging from, e.g., repressed genes, to active enhancers, or regions of heterochromatin. In the context of the International Human Epigenome Consortium [233] (IHEC), this core set comprises six histone marks that have a canonical functional interpretation usually restricted to individual marks or to combinations of at most two (Table 2.1). This set of core marks (or a subset thereof) represents the chromatin marking relevant in all remaining chapters of this thesis.

While most of the functional annotation of the individual histone marks is straightforward, e.g., H3K4me3 marking active promoter regions, others such as so-called "bivalent domains" or "poised" promoter or enhancer states demand some clarification. Initially, the term "bivalent domain" was introduced by Bernstein et al. to describe the observation that genes encoding developmentally important TFs show a distinct promoter chromatin signature in embryonic stem cells [22]. This chromatin signature consists of the two marks H3K4me3 and H3K27me3, and genes in this poised state show a low expression level despite the presence of H3K4me3 at the promoter region (cf. Table 2.1). Bernstein et al. hypothesized that these genes are kept in a state that allows for rapid activation upon suitable signaling. The existence of bivalent domains was also confirmed by Rugg-Gunn et al., who additionally described bivalent domains marked by the combination H3K4me3 and H3K9me3 [207]. Other studies extended the investigation of similar patterns of chromatin marking to enhancer elements, which established the notion of poised (or sometimes also bivalent) enhancers. In the case of poised enhancers, the chromatin signature is not reported consistently, either only referring to H3K4me1 and no H3K27ac marking [52], or to both H3K4me1 and H3K27me3 marking without H3K27ac being present at the enhancer element [197, 265]; in both cases, the presence of H3K27ac was reported to indicate an active state of the enhancer. Considerations involving two (or more) individual histone marks should always be taken with a grain of salt since the common way of identifying such bivalent domains amounts to overlaying the data of the individual histone marks and identifying co-occurrences. Customized assays that probe the same biological material sequentially for different histone marks ("reChIP") showed that the common naïve way of identifying bivalent domains lacks sensitivity and thus results in an incomplete characterization of bivalent domains [126]. Another aspect complicating the functional interpretation of histone marks lies in the symmetry of the histone core. The canonical way of looking at histone marks would suggest that, e.g., lysine 4 is either mono- or trimethylated (cf. Table 2.1), but not both because it is the same residue. This view neglects that there are two H3 N-terminal tails per nucleosome that can be asymmetrically modified [251]. In summary, the canonical functional interpretation of a limited set of core marks has been shown to be valuable to characterize the histone component of the epigenetic landscape in many important regulatory contexts.

The above sections summarize the importance of nucleosomes and of histone marks as components of various regulatory processes and highlight the flexibility inherent to many of these processes. It is thus a natural question to ask how this regulatory program is safely transmitted in its entirety from one generation to the next.

2.1.1.5 Epigenetic inheritance: are histones epigenetic factors?

Classifying (modified) histones as core components of the epigenome can be questioned on the ground of their unclear heritability. Depending on the emphasis of the heritable factor and its meaning, e.g., just across cell divisions ("mitotic inheritance") or between generations ("meiotic inheritance"), histones may fall short to meet definitions of epigenetics that emphasize the heritability of epigenetic factors (cf. footnote on page 4, Section 1.2) [55, 193, 194]. As exemplified in Sections 2.1.1.2 and 2.1.1.3, nucleosome positions are not necessarily fixed depending on the local chromatin context or on cellular processes like DNA replication and transcription. Chromatin remodeling enzymes may move or disassemble nucleosomes, which already creates the difficulty

Table 2.1: Core histone marks as defined within IHEC

Histone mark(s)	Canonical functional annotation
H3K27ac	active enhancer [52], active promoter [255]
H3K27me3	polycomb-mediated silencing [12, 221]
H3K36me3	active transcription, elongation [12]
H3K4me1	(poised) enhancer [52, 101]
H3K4me3	active promoter [12]
H3K9me3	heterochromatin [12]
H3K4me3 & H3K27me3	bivalent promoters [22, 252]
H3K4me1 & H3K27me3	bivalent enhancers [197]

of identifying the nucleosome that is assumed to carry the heritable information; a problem that is essentially absent in the world of DNA methylation[1]. This consideration also immediately hints at the other central problem: all histone proteins of the same type or variant are identical from a biochemical point of view. Hence, it seems compelling that chromatin modifying or remodeling enzymes need additional information — not represented by the histones themselves — to target the correct histones in the appropriate genomic context.

A more realistic view could be to consider the local chromatin state as the heritable epigenetic information, which would only require propagation of a "sufficient" chromatin signal to allow for re-establishment of the correct regulatory chromatin state. In the case of DNA replication, the "random model" and the "semi-conservative model" of histone propagation have been developed to describe this process [156]. Both models assume that nucleosome disassembly prior to DNA replication is followed by reassembly of new nucleosomes that consist, at least partially, of old histone H3-H4 components; these components are either distributed randomly or, as dimers or tetramers, are deposited in roughly equal proportions to the newly synthesized strand. The latter model may sound attractive in its simplicity, but it would require a process precisely regulating histone deposition in the wake of DNA replication. In the random model, however, the chromatin signal may be diluted and the transmission of the epigenetic state would thus be imprecise, but potentially sufficient. Besides this more or less direct transfer of the chromatin marking, other reports in the literature describe a propagation of the histone marks H3K4me3 and H3K27me3 during DNA replication via complete *de novo* marking by the respective HMEs [187]. Though the study by Petruk et al. [187] was conducted in fly embryos and is not directly applicable to vertebrate species such as human, it highlights the important role of HMEs as part of the (epigenetic)

[1] It is common to refer to individual methylation sites either by their genomic position (sequencing-based assays), or by an identifier that can be mapped back to a genomic position (array-based assays); personal communication with Michael Scherer.

inheritance process across cell divisions.
Elucidating the mechanisms that regulate epigenetic inheritance, especially on the chromatin level, is an ongoing endeavor, but the existing evidence seems to suggest that the combination of histone marks, histone modifying and chromatin remodeling enzymes is essential to stably propagate the cellular phenotype [38, 156]. Based on this more holistic view, histone marks are considered epigenetic factors in this thesis.

2.1.2 Non-histone proteins

Besides histone proteins, the protein family of TFs is a key player in gene regulation, which also includes the already mentioned HMEs, HBDEs and chromatin remodelers. By convention, TFs are usually not considered epigenetic, but as exemplified in Section 2.1.1.5, strict definitions of what an epigenetic factor is and is not may be premature given the pace of discoveries in the field of epigenetics. Upon activation, TFs bind to regulatory elements such as enhancers or promoters, thereby promoting access to the DNA and facilitating the recruitment of other factors, especially in the context of transcriptional initiation [125, 178]. It was already pointed out that TFs belonging to the group of HME can alter the local chromatin state, e.g., the KAT EP300. It is thus not unexpected that the presence of certain histone marks can be accurately predicted from TF binding data [21]. This relationship has also led to the realization that, within certain limits, TF binding data and histone marks are statistically redundant for the predictive modeling of gene expression [34]. Similarly, the identification of regulatory elements is also possible in a tissue-specific manner using TF binding data, e.g., using EP300 to locate and to measure the activity of enhancer elements [250]. While the notion of master TFs exists (see, e.g., Whyte et al. [256]), there is no consensus on a set of TFs to be measured routinely in analogy to the core set of histone marks (Table 2.1).

2.1.3 DNA methylation

If not indicated otherwise, the following information is based on Jones [113].

DNA methylation is the chemical modification of cytosine nucleotides in the DNA via the addition of a methyl group to the carbon at position 5. This modification is most commonly observed in the context of CpG dinucleotides. DNA methylation is commonly described as a broadly repressive epigenetic mark. Mammalian genomes usually show a high methylation level, which is presumed to be a silencing mechanism for genomic regions that must not interfere with regulatory processes. A notable exception are CpG-dense regions, so-called CpG islands (CGIs), that are often devoid of methylation and commonly occur close to regulatory regions such as promoters. Numerous studies have elucidated various patterns of DNA methylation depending on the regulatory context. For example, DNA methylation in gene bodies of expressed genes is usually high, low in CpG-rich promoters of expressed genes and variable in enhancer regions due to, e.g., the interplay with TFs. If a promoter region becomes methylated, this is usually a sign of long-term stable silencing of the respective gene, which represents an epigenetic silencing signal distinct from the histone-based bivalent domains described in Section 2.1.1.4. As opposed to histone modifications

(Section 2.1.1.5), the process by which the methylation status is copied after DNA replication is known in (more) detail [224]. Specialized enzymes, so-called maintenance methyltransferases, reestablish the methylation status on the newly synthesized DNA strand, which explains how DNA methylation status can be stably inherited during cell division. The known path of inheritance combined with the comparatively easy readout of DNA methylation status — cytosines can either be methylated or not — offer a plausible explanation why methylome analysis is often the central component of epigenome studies.

2.2 High-Throughput Assays for Chromatin Profiling

The following sections summarize a subset of current technologies that afford efficient profiling of the chromatin landscape in a genome-wide manner. The descriptions focus on the first step of the experiment, i.e., linking the epigenetic information to the local sequence context to enable localization of the epigenetic signal by subsequent HTS and short–read mapping to the genome reference. Technical details of sequencing library preparation and the biochemistry of sequencing will not be discussed (e.g., see the review by Metzker [163]). The selection of technologies is biased toward protocols that were used within DEEP.

2.2.1 Determining nucleosome positions

The observation that DNA exists basically in two activity states — open and accessible to the cellular machinery or closed — is the key to enzyme-based assays such as micrococcal nuclease digestion followed by high-throughput sequencing (MNase-seq) that generate genome-wide maps of nucleosome positions [12, 219]. During the enzymatic digestion process in MNase-seq, a micrococcal nuclease cuts DNA only at accessible sites, and with a strong preference for linker regions between nucleosomes; locations occupied by nucleosomes or by transcription factors are protected from the enzymatic digestion process. The aim of the digestion process is to generate predominantly mononucleosomal fragments with a length of around 150 bp. These fragments are then size selected via gel electrophoresis and prepared for HTS. Although MNase-seq is an established protocol for determining nucleosome positions, the reliance on an enzymatic digestion process bears the risk of introducing a bias if the digestion enzyme does not cleave the DNA in an unbiased manner [43].

A recently developed protocol, nucleosome occupancy and methylome sequencing (NOMe-seq) [121], which was also extensively used in DEEP, involves the chemical modification of DNA via a methyltransferase enzyme. The key point is that this methyltransferase methylates cytosines in GpC — but not in CpG — sequence contexts, which distinguishes this artificially introduced DNA methylation from “canonical” DNA methylation (see Section 2.1.3). Since the GpC methyltransferase can only modify accessible nucleotides, i.e., nucleotides that are not blocked by a nucleosome, NOMe-seq provides information about nucleosome position/occupancy as well as methylation status for the same DNA molecule. However, because of the sequence-dependence, regions with low GpC dinucleotide density can hardly be interrogated using this approach.

2.2.2 Interrogating histone variants and histone modifications

The arguably most widely used method for characterizing chromatin is chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) [12, 165]. The ChIP-seq protocol is usually adapted to cell type, experimental conditions, and other factors. Hence, the following description is just a generic outline of the procedure [124, 218].

To probe for histone variants in the nucleosome core or for histone marks on the histone tails, the histone core proteins are chemically cross-linked to the DNA with formaldehyde. Next, the DNA is fragmented either by means of enzymatic digestion (MNase treatment as above) or by sonication; the latter was the method of choice in DEEP. The resulting fragments, ideally of the length of the DNA of one to two nucleosomes, are then immunoprecipitated with an antibody specific for the target of interest, e.g., the histone mark H3K4me3. This step reduces the set of all DNA fragments to those that are linked to a nucleosome with the chromatin signature of interest. These DNA fragments are then used for library preparation and undergo HTS, which translates the epigenetic signal into machine-readable form. It is strongly recommended [138] to also sequence a so-called Input[2] control, which is a sample taken after DNA fragmentation but before immunoprecipitation. This control is supposed to help correcting for sample-inherent biases and needed to accurately quantify sites of histone signal enrichment (see Section 2.4).

ChIP-seq is also the technology of choice to determine TF binding sites. For this scenario, there is an adaptation of the protocol, termed ChIP-exo, that uses an exonuclease enzyme to trim the ends of the DNA cross-linked to the bound protein, thereby increasing the spatial resolution and thus enabling a more precise determination of the binding site [203]. Enhancements of the (histone) ChIP-seq protocol have been reported in terms of lowering customization needs across cell types and conditions [10], but a substantial amount of effort was — and still is — dedicated to increasing the analytical value of ChIP-seq studies by reducing the necessary amount of DNA starting material, e.g., to enable the ChIP-seq analysis of rare cell populations.

2.2.3 Toward single-cell analysis: droplet-based assay

The recent trend toward single-cell analysis in the field of NGS and epigenomics [222] has been difficult to follow for ChIP-seq analysis. The standard laboratory protocol outlined above requires a population of cells (a “bulk” sample) to collect enough DNA as starting material because certain sample preparation steps lead to a loss of material. Subsequently, artifacts inherent to ChIP-seq may render it impossible to distinguish noise from signal in the final dataset if the initial quantity of DNA was too low (see Section 2.3.1 below). While several attempts have been made to reduce the number of cells from millions to thousands [10, 33, 88, 218] or to adapt protocols to lower input requirements [213], no single-cell ChIP-seq assay has been reported in the literature so far. At the time of writing, the consensus in the ChIP-seq community seems to be that standard protocols for bulk assays will not allow downscaling of ChIP-seq analysis to the single-cell level [170].

Rotem et al. [206] reported a different approach to single-cell chromatin analysis by adapting the

[2] To distinguish the Input control experiment from any generic “input”, the capitalized spelling is commonly employed; this convention is used throughout this thesis.

technology of droplet-based sequencing (Drop-seq): Drop-seq is a massively parallel single-cell assay originally developed for transcriptome studies that combines microfluidics and DNA barcoding [153]. In the Drop-seq assay, a single cell (= the biological analyte) is encapsulated together with a barcoding bead in a droplet. Inside the droplet, the biological analyte is tagged with a unique sequence barcode via a biochemical reaction. After the barcoding step, the droplets can be dissolved and the single cells combined for subsequent sequencing; the unique barcode ensures that the resulting raw reads can be sorted according to their sample of origin. Drop-seq combined with ChIP enabled Rotem et al. to generate single-cell chromatin profiles for two histone modifications (H3K4me2, H3K4me3) in several murine cell types. However, the reported read coverage is sparse (roughly 1,000 unique reads per single cell) and hence necessitated to assay thousands of individual cells to obtain reliable profiles. This proof of concept study provided evidence that single-cell variability in the regulatory (chromatin) landscape can offer insights, e.g., into early priming in embryonic stem cells, which enables scientists to better understand paths of cellular differentiation. The authors noted nevertheless that the successful application of their method and meaningful interpretation of their results

> “relies on the existence of a coherent chromatin state in a sufficient number of sampled cells. Power to distinguish such subpopulations thus benefits from sampling large numbers of cells and from the high throughput of microfluidics systems.”

In summary, current technology does not allow for exploring the chromatin landscape at single-cell resolution with the same efficiency as for DNA methylation or transcriptome assays [75, 109]. If there is no technological breakthrough in the next few years, it is conceivable that the role of bioinformatics has to shift to decomposition methods for the bulk ChIP-seq signal, either to characterize cellular subpopulations directly or, at least, to complement other single-cell assays with a more fine-grained map of the chromatin landscape.

2.3 Sources of Noise and Validation Strategies

2.3.1 Sources of noise

The potential causes for signal artifacts in histone data are multifaceted and some of the most prevalent ones are summarized here. A core problem of the ChIP-seq technology is the requirement for large amounts of DNA as starting material. This implies that each histone dataset is a snapshot of a whole population of cells and thus shows at least biological variation, e.g., due to differences in chromatin structure or cellular state. The reliance on antibodies demands for rigorous testing and quality control, but nevertheless, antibodies have varying binding affinities and specificity, which can result in fluctuations of the histone signal that are not of biological interest [66, 138]. The enrichment of genomic loci by means of the immunoprecipitation step also implies that a certain amount of duplicated reads is to be expected. However, since the ChIP-seq protocol also involves a sequence amplification step, the source of any duplicated read is potentially ambiguous [42, 98][3].

[3] The ChIP-seq processing pipeline in DEEP was initially configured to keep all reads, and later changed to remove all reads marked as duplicates.

Regions of open chromatin and in particular highly expressed genes, have also been described as showing biased read counts and thus giving rise to false positive signals; in the context of ChIP-seq analysis to determine TF binding sites, these regions were denoted as “hyper-ChIP-able” [42, 239]. Besides chromatin characteristics, the underlying genomic sequence can also affect the signal depending on, e.g., local repeat density and thus varying read mapping efficiency; as an example, this is relevant for the heterochromatin mark H3K9me3 because heterochromatin formation often occurs in repeat-rich sequence contexts [23]. Genomic regions that are known to attract a large number of reads during read mapping are commonly referred to as “blacklist regions”, and reads overlapping these regions should be removed before any downstream processing to obtain a more realistic histone signal [11].

Since artifacts caused by the underlying biology can either be attenuated or exacerbated by details of the experimental protocol and of bioinformatic processing pipelines, large epigenome mapping consortia try to standardize sample handling and processing as far as possible ([11, 138] and Chapter 3). Despite these efforts, ChIP-seq data still present a substantial challenge for bioinformatics, in particular when samples from various research consortia should be integrated and analyzed together, which is the current objective of IHEC’s Integrative Analysis Working Group[4]. Whether or not such international initiatives will result in a generally accepted and implemented way of processing ChIP-seq data is unclear, but it seems unquestionable that early considerations [12] stating that

> “[s]ince the ChIP-Seq method is analogous to direct counting of the molecules in the ChIP DNA samples, it requires minimal normalization. The number of tags detected for a particular nucleosome is directly proportional to the modification level of that nucleosome.”

turned out to be somewhat optimistic.

2.3.2 Experimental validation

The above described various sources of technical and biological noise raise the question how the quality of a ChIP-seq experiment can be assessed. The current gold standard for validating the result of a ChIP-seq experiment is a quantitative polymerase chain reaction (qPCR) of known positive and, ideally, negative loci [11, 138]. It seems understandable that limited throughput and the necessary primer design prevent qPCR from being used to create genome-wide reference catalogs even for a single histone mark in a single condition. As a consequence, studies that include experimentally verified validation sites are limited in their scope (e.g., the software benchmark study by Micsinai et al. [164] reports a total of 297 qPCR validation sites) when compared to the actual numbers of specifically modified histones in a vertebrate genome (~10^4 – 10^5 depending on measured histone mark and experimental conditions). While bioinformatic methods can never provide proof of a successful ChIP-seq experiment, it is an accepted standard to gather evidence of success as follows: (i) if the histone mark has already been measured in a sufficiently similar biological sample, the enrichment profiles should reflect the biological similarity; (ii) if the histone

[4] ihec-epigenomes.org/about/workgroups/integrative-analysis

mark is well-characterized and known to correlate with certain genomic elements like transcription start sites or enhancers, the histone signal overlaid with the respective annotations should show the expected enrichment profile [138]. The latter *in silico* options increase in their value over time as more and more ChIP-seq experiments become publicly available and can be used for comparative quality assessment strategies.

2.4 Basic Data Processing

The background information presented so far has introduced chromatin biology as well as technologies that enable scientists to investigate certain aspects of chromatin in a high-throughput manner. This section is concerned with modeling and basic *in silico* analysis of the collected biological information. Since the connecting step between sequencing and modeling, i.e., short read mapping, is not specific to chromatin analysis, it will not be detailed here (see, e.g., Fonseca et al. [79], Trapnell and Salzberg [243] as a starting point). In other words, it is assumed that the bioinformatics analysis starts with millions of short reads aligned to the reference genome, each read annotated with its genomic coordinates and a mapping quality score. This mapping quality score can be thought of as quantifying how unique a given alignment of a read is relative to all other possible alignments of that read [141]. To provide some frame of reference, the current IHEC recommendations for ChIP-seq quality control only consider reads with a mapping quality of at least five[5], whereas, e.g., the NIH Roadmap Epigenomics Mapping Consortium (REMC) set a more conservative threshold of at least 30 for some of their ChIP-seq analyses [205]. It should be pointed out that there is no commonly accepted minimum mapping quality for reads to be kept for downstream analysis of ChIP-seq data (besides a mapping quality larger than zero). Hence, within this thesis and as far as relevant and controllable, any value that at least fulfills the IHEC recommendations is considered acceptable.

2.4.1 Data reduction: genome-wide histone signal

A standard way of representing aligned reads along the genome is in the form of so-called signal or coverage tracks. The basic idea is to convert the read alignment information into a read count per bp or per genomic bin of fixed size, say, 25 bp. There is no common standard for generating signal tracks, e.g., concerning quality filtering or signal smoothing. An example for one of the few more widely applied processing steps is the normalization of the read counts to an average genomic coverage of one by appropriate scaling (“1x coverage normalization”) [226]. The genomic coverage is computed as

$$\text{genome coverage} = \frac{\text{\# aligned reads} \cdot \text{read length}}{\text{effective genome size}} \tag{2.1}$$

The effective size of a genome refers roughly to the part of the genome to which short reads can be aligned, i.e., it is a parameter dependent on read length and has some other pitfalls that are not of

[5] github.com/IHEC/ihec-assay-standards

interest here[6]. Since signal coverage tracks are usually at least one order of magnitude smaller than the original alignment file, they are a default output of many epigenome processing pipelines to facilitate further downstream analyses and data sharing (see Chapter 3). For example, signal tracks make it straightforward to quickly examine the strength of the individual histone marks in regions of interest, e.g., by taking the mean of the signal values in gene promoters. For many applications, this is already a sufficient characterization of the local genomic activity level (cf. Table 2.1, see Chapter 5).

2.4.2 Data modeling: detecting sites of enrichment

Whereas genome-wide signal tracks provide a convenient abstraction for visualizing and processing the read information of a ChIP-seq experiment, a substantial part of the biological information can be captured in a much more compact way. To this end, so-called peak calling software is used to locate genomic sites (“signal peaks”) that show enrichment of reads relative to a suitably chosen background model. These sites are then assumed to indicate positions of nucleosomes that have the histone marking of interest, e.g., H3K4me3 or H3K36me3. Within the context of this thesis, ChIP-seq background always refers to the read distribution observed in the Input control experiment. Software tools for the task of peak calling exist in such a large variety that examples named in the following were selected only for referential purposes and are by no means exhaustive (for an overview, see, e.g., [164, 242]).

In essence, two modeling decisions are implemented in most peak calling tools: how to define an appropriate size for the window of enrichment and how to model the read distribution in order to attribute the notion of significant enrichment relative to the background for the identified peak. Concerning the first point, there is a plausibility argument that due to the length of the nucleosomal DNA plus a stretch of linker DNA, the minimal window size should be around 200 bp. In practice, the window size is often estimated from the data and realistic values are in the range of approximately 140–180 bp. In principle, a strongly positioned nucleosome with high occupancy, i.e., a nucleosome present in almost all cells in the bulk sample, could be detected with a signal peak of that size, which is commonly described as a sharp or narrow peak. The complementary situation of diffuse enrichment covering several neighboring windows with difficult to detect boundaries is called a broad peak. It is customary to classify histone marks by their predominant enrichment profile, e.g., H3K4m3 and H3K27ac are canonical narrow marks and H3K36me3 and H3K27me3 are considered broad marks. This is, however, a somewhat artificial classification scheme, because such distinctions do not exist *in vivo* and varying peak shapes for different histone marks have been observed [18]. The determined window size is then used in a binning (e.g., histoneHMM [100], specialized for broad marks) or sliding window approach (e.g., MACS [268], general purpose) to scan the genome for enriched sites. To determine whether the local read count is statistically significantly higher than expected, the second modeling decision of how to model read counts needs to be addressed. Again, given the variety of existing peak callers, the description of this modeling step is limited here to a common variant (an overview of other approaches for several peak callers can be found in Thomas et al. [242]). Moreover, peak

[6] deeptools.readthedocs.io/en/latest/content/feature/effectiveGenomeSize.html

callers such as MACS [268] offer several different ways of parameterizing the background read distribution. To simplify matters, it will be assumed that the Input control experiment is used to globally model the background.

The next paragraph follows Hilbe [104] in argument and notation if not stated otherwise.

A standard way of modeling read counts is to assume that the counts X follow a Poisson distribution with mean parameter μ: $X \sim Poisson(\mu)$, e.g., in MACS [268]. In that case, the probability of observing k reads within a window is given as

$$Pr(X = k) = e^{-\mu} \cdot \frac{\mu^k}{k!} \tag{2.2}$$

By estimating the parameter μ from the Input control, i.e., the assumed null distribution of read counts, it is possible to compute a p-value for the enrichment as

$$Pr(X \geq k) = 1 - e^{-\mu_{\text{Input}}} \sum_{x=0}^{k-1} \frac{\mu_{\text{Input}}^x}{x!} \tag{2.3}$$

The Poisson assumption works reasonably well in practice, in particular for narrow peak calling. However, there is limited flexibility in the estimation of mean and variance with the single parameter μ, which leads to a lack of fit for data where the assumed equality of mean and variance does not hold. This situation is called extradispersion, which is commonly encountered in the form of overdispersed data, i.e., the parameter μ underestimates the true variance in the data (“excess variance”). Since overdispersion is frequently observed in ChIP-seq experiments, modeling the read counts X as a negative binomial is a common alternative strategy: $X \sim NB(\mu, \alpha)$, e.g., in ZINBA [201] or histoneHMM [100]. Here, the additional parameter α is the dispersion parameter. The effect of the dispersion parameter can be understood by examining the variance σ^2 of the negative binomial distribution

$$\sigma^2 = \mu \cdot (1 + \alpha \cdot \mu) \tag{2.4}$$

The parameter μ is the mean count value as above, and one can see that an increase in the mean leads to an increase in the variance, and the variance is never smaller than the mean. In the special case of $\alpha = 0$, this reduces to the Poisson model with the variance equal to the mean. Under this parameterization, the probability of observing k reads is given by

$$Pr(X = k) = \binom{k + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{\alpha\mu}{1 + \alpha\mu}\right)^k \left(\frac{1}{1 + \alpha\mu}\right)^{\frac{1}{\alpha}} \tag{2.5}$$

Since the use of a negative binomial necessitates the introduction of the additional overdispersion parameter α, the model estimation is more involved compared to the regular Poisson model and will not be elaborated here (a complete technical description for modeling overdispersed read counts can be found, e.g., in [147]). Whether or not this additional complexity is actually needed, is not easy to answer. Count data can appear to be overdispersed but may in fact not be, and a few simple adjustments, e.g., removing outliers or dealing with non-randomly missing values, could result in (transformed) count data that can be fit well with a simple Poisson model (see Tables 2.2 and 2.3 in Hilbe [104]).

The last adaptation of read count modeling presented here attempts to adjust for the following ambiguity: assume that the ChIP-seq experiment is supposed to detect nucleosomes with H3K4me3 histone marking. When scanning the genome for sites of enrichment, a peak calling algorithm inevitably also finds many sites with a read count of zero. Note that the situation of $X = 0$ can be captured with both the Poisson and the negative binomial model and is then expected to be encountered with a certain probability according to the model. A read count of zero could occur at the location of a nucleosome that does not have the required histone marking, but the marking could in principle be acquired at some point; here, the zero value has a biologically meaningful interpretation. If, on the other hand, the zero read count is caused by, e.g., a technical problem such as insufficient sequencing depth, the zero value is biologically meaningless. This consideration suggests that the number of zero counts can be excessive relative to the expectation based on the chosen count model. This situation is referred to as "zero excess" and can be handled, e.g., by adapting the count model for zero inflation. Under zero inflation, the read counts are assumed to have been generated by two distinct processes, one just producing (excess) zeros and the other one producing the full range of counts including "meaningful" zeros. Zero-inflated models are implemented in only a few peak calling algorithms, e.g., in ZINBA [201] and Zerone [53]. This may be due to the observation that the largest improvements in model fit are realized by switching from a Poisson to a negative binomial model for overdispersed data, and additional improvements by explicitly modeling excess zeros are only minor [57].

It should be pointed out that despite all considerations about appropriate modeling of (ChIP-seq) read count data, no community-wide accepted way of modeling ChIP-seq read counts and performing the subsequent peak calling has been established; a goal that might also be hard to achieve given the numerous sources of noise (see Section 2.3.1). Consequently, any study that relies on peak calling as an intermediate step bears the risk of calibrating downstream analyses to the peak calling software, which could negatively affect reproducibility of the results (see Chapter 3). To illustrate that point by example, Table 2.2 lists genomic coverages for all peak calls of several peak calling tools applied to a triplicated hepatocyte control sample from DEEP. The two tools MACS2 [268] and histoneHMM [100] (only broad marks) are part of the official DEEP processing pipeline for ChIP-seq data. The peak caller Zerone [53] was added for comparison and produces a single list of peak calls for all replicates combined (the donor number is given as "NN" in these cases). This exemplary case shows substantial variation in genomic coverage of the called peaks, suggesting that common downstream analyses that rely, e.g., on assessment of peak overlaps with other functional annotation data may give different results when switching to another peak calling algorithm. Although one may argue that different output from different tools employing different modeling assumptions about the data is no surprising observation, this would miss the point. Because there is no generally accepted way of combining peak calls across any number of biological replicates and from any number of different tools, and genome-wide gold standards of the true biological signal do not exist (see Section 2.3.2), the situation illustrated in Table 2.2 represents one of the persistently challenging problems in ChIP-seq bioinformatics. Hence, any

Table 2.2: Example of differences in genomic coverage for peak calls produced by different tools. The three selected samples are biological replicates of hepatocytes (healthy controls) from three male donors. The samples were sequenced, aligned and analyzed (except for tool Zerone) in the context of the DEEP consortium.

Donor	Mark	Profile	Coverage (Mbp)	Software
Hm09	H3K27ac	narrow	173.81	MACS2
Hm16	H3K27ac	narrow	260.18	MACS2
Hm25	H3K27ac	narrow	169.29	MACS2
HmNN	H3K27ac	narrow	140.20	Zerone
Hm09	H3K27me3	broad	550.32	MACS2
Hm09	H3K27me3	broad	1430.98	histoneHMM
Hm16	H3K27me3	broad	749.38	MACS2
Hm16	H3K27me3	broad	1525.89	histoneHMM
Hm25	H3K27me3	broad	588.34	MACS2
Hm25	H3K27me3	broad	1457.07	histoneHMM
HmNN	H3K27me3	broad	847.94	Zerone
Hm09	H3K36me3	broad	430.94	MACS2
Hm09	H3K36me3	broad	971.12	histoneHMM
Hm16	H3K36me3	broad	456.20	MACS2
Hm16	H3K36me3	broad	980.06	histoneHMM
Hm25	H3K36me3	broad	423.73	MACS2
Hm25	H3K36me3	broad	963.98	histoneHMM
HmNN	H3K36me3	broad	307.20	Zerone
Hm09	H3K4me3	narrow	64.84	MACS2
Hm16	H3K4me3	narrow	69.02	MACS2
Hm25	H3K4me3	narrow	64.68	MACS2
HmNN	H3K4me3	narrow	42.88	Zerone

study-specific solution to the problem of how to handle peak calls in downstream analysis bears the risk of not resulting in generalizable conclusions.

CHAPTER 3

Reproducibility in Computational Research

Lead-in This chapter is concerned with possible solution strategies for the problem of lacking reproducibility of scientific results — commonly referred to as the "reproducibility crisis". The following material represents the state-of-the-art approach toward improving reproducibility in the context of the collaborative research initiative DEEP. Section 3.1 provides the necessary background in a broad, i.e., non-DEEP specific sense, whereas the remaining sections are written from the DEEP perspective. This chapter concludes with Section 3.6 as a detailed outlook on potential future developments to support reproducible research in computational epigenetics, which, for reasons of chapter coherence, will not be summarized as part of Chapter 6 (Perspectives).

The work presented in this chapter is an extended version of the manuscript Ebert et al. [64] (for publication details and author contributions, see Appendix E.1.1).

3.1 Background - What is Science?

A necessary precondition to any discussion on the topic of scientific reproducibility is to define science itself. The body of work on this subject is vast and the different philosophical considerations are beyond the scope of this work. A possible and concise definition of science is given by Edward Osborne Wilson [257]:

> "Science [...] is the organized, systematic enterprise that gathers knowledge about the world and condenses the knowledge into testable laws and principles."
> (Chapter 4 "The Natural Sciences")

It is instructive to also highlight two of the main characteristics of science:

> "The diagnostic features of science that distinguish it from pseudoscience are first, repeatability: The same phenomenon is sought again, preferably by independent investigation, and the interpretation given to it is confirmed or discarded by means of novel analysis and experimentation. Second, economy: Scientists attempt to abstract the information into the form that is both simplest and aesthetically most pleasing —

> the combination called elegance — while yielding the largest amount of information with the least amount of effort.”
>
> (Chapter 4 “The Natural Sciences”)

Following Wilson’s arguments, creating conditions that support the independent re-investigation of natural phenomena and the presentation of abstracted information in a simple and accessible way are two pillars of science. While the latter is open to interpretation, in particular concerning the question what the “aesthetically most pleasing” form of abstraction is, the importance of independent testing has also been discussed in many epistemological treatises. To give an example, in his work *The Logic of Scientific Discovery* [190], Karl Popper argues that

> “a positive decision can only temporarily support the theory, for subsequent negative decisions may always overthrow it. So long as theory withstands detailed and severe tests and is not superseded by another theory in the course of scientific progress, we may say that it has ’proved its mettle’ or that it is ’corroborated’ by past experience.”
>
> (Section I.1.3 “Deductive Testing of Theories”)

> “Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested — in principle — by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated ‘coincidence’, but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable.”
>
> (Section I.1.8 “Scientific Objectivity and Subjective Conviction”)

It is the premise of this chapter that these theoretical considerations are accepted as vital for scientific progress by the majority of scientists. It is, however, not obvious how these principles should be translated into practical tools assisting researchers in their daily routine. The work presented here is a proposal of concepts and tools that were developed with the objective of bridging the gap between theory and daily practice (Section 3.2).

3.1.1 Disentangling replicability, reproducibility and repeatability

In the previous section, different terms are used interchangeably that all refer to the concept of scientific reproducibility. In the area of computational research, it is of particular importance to distinguish between these terms to avoid common misconceptions; following the definitions of Drummond [59], “replicability” will be defined as rerunning code to produce the same output again. This mere replication naturally includes all potential flaws in the results, e.g., due to programming errors in the software code. Successful replication thus cannot be seen as evidence supporting the original claims. In particular, replicability does not imply that the reported results would hold up in an independent test. On the other hand, lack of replicability implies lack of reproducibility.

“Reproducibility” refers to an independent test or attempt to reproduce results using a different

and yet suitable approach that should allow drawing the same conclusions from the same data. In computational research, the implementation of various distinct algorithms or models all of which aiming at the same goal represents a setting where reproducibility can be easily tested. Of course, this example neglects the data generation part, i.e., a noisy input dataset may generally render any study irreproducible. Since the scope of this chapter is on the computational side, potential reproducibility issues caused in the wet lab are not addressed. More precisely, computational reproducibility only supports the original claims about the information contained in and extractable from a published dataset. Assuming a fully deterministic analysis, one may argue that a reimplementation of sufficiently similar methods would always result in the same output, which should not be taken as evidence of reproducibility. Although a formal definition of "sufficiently similar methods" is probably hard to conceive, such criticism makes nevertheless intuitive sense; it emphasizes that "reproducibility requires change" [59]: an independent test has to aim at the same goals, but with different means.
The term "repeatability" appears to be defined with most stringency and usually refers to the repetition of measurements within the same experimental setup, e.g., to assess the precision and reliability of one or several detection methods [26]; it is thus not relevant for this chapter.

The notion of reproducibility as defined above will be used throughout the rest of this chapter. However, there is one caveat due to reasons of practicality: because reproducibility requires an independent outside perspective, it is difficult for the same individual to truly work reproducibly. As a consequence, this chapter uses a "working definition" of reproducibility that requires replicability and aims at transparency of workflows. It is assumed that reproducing any computationally generated result is feasible under these conditions. In essence, transparency requires a thorough documentation of computational analysis steps and unambiguous ways of identifying the correct datasets for the analysis; given that, a computational analysis as a whole is testable in the sense that it should be possible to reach the same conclusions. The exact identification of datasets is a manageable — if not trivial — task, whereas the concise documentation of an analysis pipeline can be a challenge that grows with the pipeline's complexity.

3.1.2 Planning for reproducibility: a tradeoff in granularity

A pivotal question for an environment that fosters reproducible research is what level of detail — or granularity — can be considered "sufficient" in terms of capturing enough information to ensure reproducibility of a body of work. In simpler terms, there does not seem to be a universal standard for reproducible research that defines what information needs to be recorded and to what level of detail. Computational work in particular presents a challenge to defining the level of required granularity: numerical simulations should be documented with exact information on the numerical libraries and, potentially, on the underlying hardware, whereas the description of a data conversion step in a computational pipeline may arguably not require such an amount of detail. The assumption in this chapter is that the above question can only be answered in a context-specific manner. This view is supported by Thain et al. [240] who state that

> "[t]he choice of granularity depends on the overhead of metadata management, storage overhead, the time overhead of submission, storage and reconstruction, and user-friendliness."

Assuming user-friendliness as crucial to foster compliant behavior among scientists implies that the granularity should be governed by ease of accessibility. The information a scientist is asked to record should be immediately available, e.g., the version of a software tool is usually easy to identify whereas the versions of all programming libraries the tool relies on may not be discoverable. The approach of favoring simplicity over completeness[1] can be beneficial as it captures less but more meaningful information. Of course, if attempts to reproduce a study fail due to incomplete documentation, a previously defined schema has to be revised to record the missing information in future studies.

Although the unclear requirements regarding record keeping in practice may seem like a potential candidate that could at least partially explain the reproducibility crisis, Section 3.1.3 summarizes in the following several recently published articles on the matter that do not seem to motivate the conclusion that more (comprehensive) guidelines on information recording can solve the reproducibility crisis.

3.1.3 Current issues with scientific reproducibility

The lack of reproducibility in biomedical research has recently been addressed by the editors of more than 30 major journals [171] and is an issue of active debate [172]. In the published editorials, the authors report on a common set of Principles and Guidelines in Reporting Preclinical Research[2] that should support reproducibility by establishing clear requirements for any manuscript to be eligible for publication. Although such an initiative driven by the *Nature Publishing Group* and by *Science* is certainly a necessary step to raise overall awareness for the reproducibility crisis, a closer look at those guidelines reveals that they are still abstract in nature; the following three examples are an excerpt from the National Institutes of Health (NIH) guidelines:

- "A section outlining the journal's policies for statistical analysis should be included in the Information for Authors, and the journal should have a mechanism to check the statistical accuracy of submissions."
- "Standards: Encourage the use of community-based standards (such as nomenclature standards and reporting standards like ARRIVE), where applicable."
- "Encourage presentation of all other data values in machine readable format in the paper or its supplementary information. Require materials sharing after publication."

These guidelines leave room for interpretation: who exactly is supposed to check the statistical accuracy and to what level of detail? How should the reporting be done if there are no community-based standards? Who checks that the authors did not overlook or ignore existing standards? What is a machine-readable format? Does that, e.g., include proprietary binary file formats?

Similar guidelines or best practices were published in numerous other articles (e.g., Peng [184],

[1] Thain et al. [240] refer to this as "capturing the mess" versus "encouraging cleanliness".

[2] nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

Sandve et al. [212], Schnell [217]) and it is reasonable to assume that constant reminders on the topic are necessary to avoid a qualitative drop in reproducibility standards. However, there is presumably a consensus in the scientific community that burdening reviewers with the practical implementation of all those guidelines is not a viable solution. It is thus an open question how a practically useful realization of these quality control measures should look like. The lack of such concrete implementations can lead to presumably widespread and often trivial errors. For example, Ziemann et al. [272] reported on a substantial number of erroneous gene symbols in the supplementary material of 704 out of 3,597 studies published in 18 different journals. The source of the error was identified as automated string format conversion by Microsoft Excel®. It should be emphasized that depositing lists of gene symbols in the form of spreadsheets is in line with the above guidelines (use a “machine readable format” and adhere to “nomenclature standards”). Published “how to work reproducibly” articles and guidelines provide a conceptual framework that outlines necessary conditions for reproducible work. Yet, these articles commonly lack practical solution strategies that support scientists in working reproducibly as part of their daily routine.

The multifaceted roles of researchers in the scientific community as authors, reviewers, editors, etc., may exacerbate a stringent separation of responsibilities. It nevertheless stands to reason that responsibilities and actions have to be defined and distributed across the scientific community to deal with the reproducibility crisis. One could assume that scientists will always work reproducibly[3], because irreproducible work is not publishable — the current reproducibility crisis proves that wrong. It is probably unrealistic that reproducibility can be guaranteed under all conditions, because this would require to eliminate, e.g., negligence, miscommunication, and human error. Therefore, practical solutions should presumably aim at managing reproducibility to bring down the number of irreproducible studies [14, 67, 169, 192].

3.2 Problem Analysis and Project Objective in the Context of DEEP

The organizational structure of DEEP is similar to that of other research consortia such as the Encyclopedia of DNA Elements [49, 231, 241] (ENCODE), REMC or the European Hematopoietic Epigenome Project [1] (BLUEPRINT). The consortium structure reflects the collaborative nature of large-scale scientific projects that rely on expertise from various groups that are not (necessarily) located on the same research campus. This implies that the consortium-related work in an institute is usually built on top of existing infrastructure and internal operating procedures. For example, the DEEP compute cluster at the Data Analysis Center (DAC) is administrated by the DAC’s Information Services and Technology (IST) department following all in-house policies, with limited exceptions to allow for access by selected external users. As a consequence, any attempt to develop and to implement a consortium-wide strategy for recording (computational) metadata for the purpose of supporting reproducibility has to take into account the existing structures at all collaborating institutes. The role of each contributing research group in DEEP can be roughly characterized using one or two of the following tasks:

[3] The caveat mentioned in Section 3.1.1 should also be taken into account here.

1) Acquisition: collecting primary samples from donors in clinics or, in case of animal data, in specialized facilities; cultivating cell lines in the wet lab
2) Translation: running assays such as ChIP-seq to translate the biological information into the (computer-readable) form of data files containing raw sequencing reads
3) Quantification: performing the low-level of bioinformatics analysis to obtain an easier to interpret quantitative representation of the biological signal; short read mapping to the reference genome and generating standard output (e.g., signal coverage tracks, see Section 2.4) are both subsumed under this point
4) Interpretation and integration: performing the high-level bioinformatics analyses, e.g., differential analysis between samples or exploratory analyses guided by specific research questions
5) Collection and storage: collecting all (digital) results for archival purposes in a form such that the research community can access and analyze the data; sensitive information that may be used to identify (human) individuals is excluded from open access portals for ethical reasons

Each of these tasks comes with a different burden of metadata complexity. For the sample acquisition, there will be one metadata record per sample at the end of the consortium's life span. Experiment and translation already result in several metadata records per sample because the minimal requirements for an IHEC reference epigenome specify nine assays for each sample[4]. Of the three remaining tasks, which are mostly the responsibility of computational research groups, only the initial quantification of raw data can be documented with a predefined metadata model, because it merely translates raw data into commonly used standard file and data formats; a task for which it is comparatively straightforward to design analysis pipelines with a more or less fixed scope. Such pipelines enable scientists to formulate expectations about reasonable metadata to be recorded for a pipeline run. As opposed to quantification analyses, downstream interpretation and integration tasks are commonly guided by research questions and, thus, analysis code may change frequently. This complicates the *a priori* definition of milestones for a study that would lend itself to capturing a coherent metadata record up to that point (with the sole exception of the final publication stage). Similarly, data collection and storage require a continuous effort within a consortium that lacks a formal definition of data relevance ("what to record?") and of data quality ("when to record?"). One major consequence of the above points is that the development of a practically useful metadata model for computational analyses should start with an emphasis on the initial quantification step. This approach has been realized in DEEP and is the primary concern of this chapter.

The requirement of reproducibility predates the era of large-scale computational research and, thus, numerous concepts and software tools aiming at documenting computational analysis for supporting reproducible computational research have been published in the past couple of decades. In the following section, several of these will be presented with a focus on contemporary software solutions that were also evaluated in terms of their potential applicability within DEEP.

[4] Bisulfite-seq, RNA-seq, ChIP-seq Input plus six histone marks; for a list of optional assays, see ihec-epigenomes.org/research/reference-epigenome-standards

3.2.1 Existing concepts and software solutions to record computational metadata

3.2.1.1 Literate programming and interactive reports

As explained in Section 3.1.1, reproducing a computational analysis requires an independent reimplementation of its concepts and underlying ideas. For any individual software tool or core algorithmic component, this suggests that the code itself needs to be described in a manner that exceeds inline comments or auto-generated documentation of library functions. This style of writing code has been put forward by Knuth [127] as “Literate Programming” with an emphasis on “explaining to *human beings* what we want a computer to do”. The idea of Literate Programming is at the core of many software packages (e.g., Perez and Granger [186], Xie [262]) that aim at simplifying the creation of so-called interactive reports. These packages create documents that mix code, textual documentation and embedded graphics with the ability to rerun the code using altered parameter settings at the click of a button[5]. For an isolated analysis with limited scope, an interactive report presumably contains a complete record of all data transformations plus the relevant code, hence enabling independent reproduction of the results. However, in a consortium setting, standardized analyses commonly consist of several chained software tools developed by different research groups. This suggests that the ideas of Literate Programming and their modern interpretation in the form of interactive reports may not be directly applicable to documenting complex analysis processes; from the consortium perspective, the inner workings of each individual software tool are less relevant than their combined setup. Restricting the documentation to the overall workflow, i.e., to the series of computational analysis steps, affords a strict separation between workflow specification and its execution, which avoids redundancy in documentation.

3.2.1.2 Software for managing computational workflows

Popular tools in the bioinformatics community that implement a workflow-centric approach are Galaxy [2, 87], Taverna [176, 259], and KNIME [24][6]. Galaxy provides a rich graphical user interface tailored to creating “clickable” workflows that do not require any programming skills to set up, provided that all tasks in the workflow can be accomplished with tools registered in the Galaxy repositories[7]. These workflows can be downloaded as textual documents (JavaScript Object Notation (JSON) serialized) and thus can be shared with collaboration partners. Of course, taking full advantage of sharing Galaxy workflows requires functional Galaxy instances at the receiving end. Although the burden of a local Galaxy setup including configuration and maintenance could be avoided by using off-the-shelf Galaxy cloud instances, analyzing sensitive patient data is problematic due to the inherent privacy issues in such a scenario. Sharing Galaxy workflows with the purpose of just informing collaboration partners about a computational

[5] It is common to refer to such interactive reports as “computable documents”. Despite being an appealing notion, it should be pointed out that the “Computable Document Format”™ (CDF) is a proprietary document specification by Wolfram Research.

[6] KNIME’s desktop application is free of charge, but support for compute clusters requires a paid license. Due to this restriction, the focus is on Galaxy and Taverna.

[7] toolshed.g2.bx.psu.edu

pipeline is similarly cumbersome: Galaxy workflow documents are meant to be imported into Galaxy and then to be displayed in a human-readable form, i.e., besides the actual workflow steps, these documents contain additional information pertaining to the layout, to resource identifiers and to Galaxy-internal attributes (see example in Appendix A.1.1). In summary, Galaxy workflows in their textual form are not meant to be read by humans and are arguably not ideal for disseminating descriptions of computational pipelines.

A popular alternative to Galaxy in the bioinformatics community is Taverna. Taverna offers similar capabilities as Galaxy, e.g., a client software with a graphical interface for creating and manipulating workflows, or specialized setups targeting cloud computing environments. Taverna's feature set comes with the same downside regarding limited human readability of workflow documents. Originally, Taverna workflows were also specified in the form of Extensible Markup Language (XML) documents, but since the release of Taverna 3, Taverna's Simple Conceptual Unified Flow Language (v2) (SCUFL2) specifies entire workflow bundles[8]. These workflow bundles are containers that, besides the workflow document, can include various metadata records such as execution profiles for different computing environments, or annotations for resource identification and data provenance tracking. Given that Taverna also emphasizes the use of web services as part of computational pipelines, it is understandable that Taverna offers a variety of ways to semantically enrich workflow bundles. Although Taverna's workflow bundles are a powerful way to develop and to document scientific workflows, they only show their full potential when imported into Taverna; here, similar considerations concerning consortium-wide setup and maintenance as for Galaxy apply.

By restricting the view on bioinformaticians, it would be feasible to select software tools for workflow development and documentation that do not offer a graphical user interface (e.g., Snakemake [131], Ruffus [90] or (GNU) Make[9]). In this scenario, it can be expected that the resulting workflow documents also contain a considerable amount of program code. Consequently, such workflow documents can only be reasonably shared with third parties that have the right set of programming skills for understanding the workflow. Of course, using workflow documents created by any of the aforementioned command line tools leads to a similarly limited usability as described above for Galaxy and Taverna. A possible way out of this "locked-in"[10] situation is provided by standards for workflow documents that are cross-platform compatible; one currently emerging standard is the Common Workflow Language (CWL)[11]. The development of CWL was initiated with the goal of establishing the *de facto* standard for specifying complex, cross-platform compatible computational workflows in data-intensive fields such as bioinformatics. If such a standard were supported by a large enough number of software tools and found acceptance

[8] There is no example of a Taverna workflow bundle included in the Appendix since it is not practical to visualize its structure and content in textual form.

[9] gnu.org/software/make: presumably the spiritual father of all modern workflow management frameworks.

[10] The term "(vendor) lock-in" is usually used for commercial or otherwise closed software; in this context, it is simply referring to the fact that the workflow descriptions of the individual tools are not cross-compatible.

[11] commonwl.org: Common Workflow Language stable specification v1.0. Recently, other new projects such as the Broad Institute's Workflow Description Language (WDL) (software.broadinstitute.org/wdl) have emerged, which indicates that several modern workflow languages are competing for becoming the standard in bioinformatics.

throughout the scientific community, this would indeed lower the need for human readable workflow documents if workflow visualization capabilities are built into commonly used software tools.

In summary, existing software solutions emphasize usability and ease of developing new workflows and, due to the lack of standards, can only ensure proper documentation of computational analysis within the boundaries of their own ecosystem. As explained in Section 3.2, existing software infrastructure has to be taken into account in a collaborative project. It is therefore questionable whether current workflow management tools offer the right set of features to fulfill the needs of several computational research groups working under the same umbrella. The obvious alternative to software that assists in developing and executing workflows, is to focus just on documenting computational workflows and to leave the implementation details at the discretion of the respective computational research group.

3.2.1.3 Code-free reporting standards

In biology and biomedicine, code-free reporting standards exist for numerous purposes, e.g., as part of the Minimum Reporting Guidelines for Biological and Biomedical Investigations (MIBBI) project [238]. At the time of writing, none of the available standards explicitly addresses *in silico* workflows, though they cover current experimental setups for high-throughput assays[12]. A related initiative called ISA-tab[13] focuses on decomposing large projects, e.g., DEEP as a whole, into smaller studies that collect and process data by running various assays. The ISA-tab framework offers templates and different tools that help generating the necessary documentation by providing convenience functions such as the incorporation of controlled vocabularies via ontologies. Since the ISA-tab framework is generic by design, it is not impossible to document computational workflows, yet the examples provided by ISA Commons strongly suggest that this is not the intended usecase for ISA-tab. Moreover, rationality dictates that metadata records following the ISA-tab specification should be used throughout the entire consortium, i.e., from wet lab to dry lab, to achieve the highest benefit from implementing such a structured approach.

3.2.2 Documenting research projects: a matter of culture

A possible conclusion resulting from the comments on software-based and code-free solutions for workflow specification and documentation is that they reflect two fundamentally different approaches toward the same goal: computational scientists aim at offloading the tedious task of record keeping onto the computer. Contemporary workflow management systems easily scale with the growing amount of data and, at least as long as best practices concerning data backup are followed, no information about analysis runs is lost. Accessing computer-generated documentation and automatically created metadata records within a heterogeneous consortium is nevertheless still a challenge, presumably due to the lack of a predominant data exchange

[12] Examples: Functional Annotation of Animal Genomes (FAANG) experiment and sample metadata specification or the Minimum Information about a high-throughput Nucleotide Sequencing Experiment (MINSEQE) reporting guidelines.

[13] isacommons.org: data model for “Investigation-Study-Assay” tab-separated data.

standard. In other words, there seems to be an emphasis on record completeness at the expense of record accessibility by collaboration partners.

At the other extreme of the spectrum are manually created documents, which seem to be more commonly used in "non-computational" research settings. These range from handwritten notes in lab notebooks[14] to more sophisticated approaches relying on supportive software and templates as exemplified above for the ISA-tab initiative. Of course, manual work does not scale and is prone to missing information due to, e.g., human error. However, the manually recorded information is more accessible by third parties[15], especially if the information exchange happens exclusively among humans.

When limiting the view to the existing software solutions, an important observation can be made: solving the problem of ensuring replicability of a body of work seems to be feasible[16]. Replicability is already a central step toward reproducibility (Section 3.1.1). However, the limitations of existing software solutions in terms of difficult deployment throughout a heterogeneous consortium, potentially restricted handling of data and metadata and unclear support for newly developed tools sparked reasonable doubt that any off-the-shelf method would be suitable to generally capture computational metadata within DEEP. Hence, when designing the computational metadata model for DEEP, the objective was to create a modest solution that would enable replication — and ideally reproduction — of all computational analysis in DEEP while striking a balance between manual documentation and automated information recording.

3.3 Solution Strategy Developed for DEEP

Taking into account the central issues identified in the problem analysis and the perceived shortcomings of existing off-the-shelf solutions for workflow documentation, the goals for the DEEP computational metadata model can be characterized with the following conceptual requirements:

1. Modesty: the computational metadata model should serve its purpose with only a limited amount of (manual) work required by computational scientists in the consortium (exceptions apply for members of the team who developed the specification). Modesty helps focusing on central problems and increases the chances of acceptance as the necessary mental effort to understand and to use the metadata model is low.
2. Iterative refinement: the different roles of the collaboration partners (Section 3.2) in a consortium are clear from the start, but the data or information contributed by them may change over time. As a consequence, the metadata model and related tracking systems need to be designed flexibly enough to allow for updates or extensions without a restart of the whole system.

[14] As done in DEEP — personal communication with members of the AG Walter, Saarland University.

[15] Ease of sharing considered as orthogonal problem here.

[16] Substantial improvements in facilitating replicability of computational analyses have been made in the past couple of years due to the steep increase in popularity of tools such as Bioconda that aim at standardizing software environment setups and also support the packaging of entire workflows as containers [92].

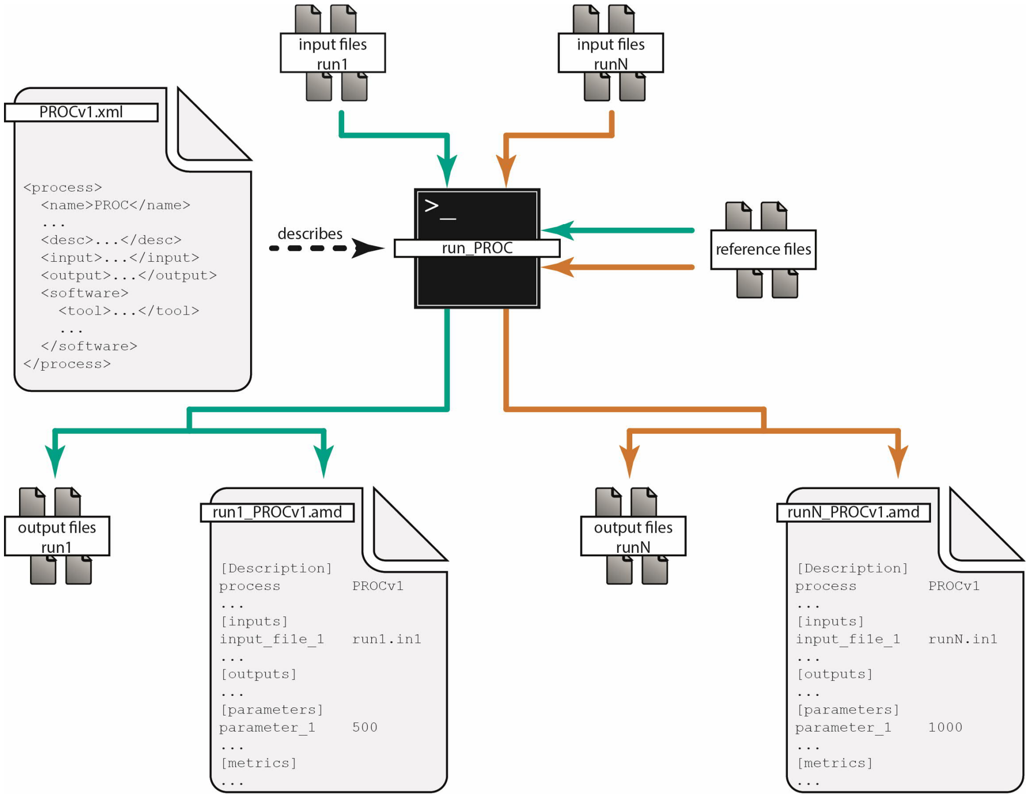
3. Human communication: access to the tracked information must be easy enough to allow for fluent communication and information exchange by all researchers involved in the project, with different levels of granularity depending on their scientific background. Although not optimal in terms of transparency and accessibility, disseminating information via e-mail is still common and it is thus desirable to have metadata available in a form that supports making use of that medium if necessary.
4. Computer communication: a research setup distributed across several institutes implies a diverse software landscape that requires simple text- or API-based data exchange if no single *de facto* data standard exists. In particular, application-specific binary data formats or popular yet proprietary file formats like Microsoft Excel® spreadsheets are not suitable for hassle-free interoperability.
5. Controlled collection: if possible, any metadata record should be validated before entering the DEEP ecosystem and erroneous data should trigger immediate feedback to the submitting user to ensure timely correction. Besides checking for errors, controlled metadata collection encompasses the task of avoiding redundancy in the records, which can be a prime source of confusion for users when querying the system for information.
6. Centralized collection: a single authority has to collect the complete body of metadata to avoid synchronization issues between collaborating institutes and to offer the same coherent information to all scientists in the consortium and in the scientific community.

All concepts developed at the DAC to record metadata of computational analyses within DEEP implement one or more of these points. However, since these concepts were designed and tested as part of DEEP itself, it was not possible to realize the last point of a single source of information, which naturally would have been hosted and maintained by the Data Collection Center (DCC) in Heidelberg. The following paragraphs provide an overview of the metadata model developed as part of DEEP. The technical implementation details are presented in Section 3.4.

3.3.1 Concept digest: process and analysis metadata

DEEP metadata for a computational analysis pipeline consist of two core components, a generic "process document" and an "analysis metadata file" (Figure 3.1). A process document provides a template-like specification of a computational analysis run, listing required input, output and reference data plus the individual command line calls to execute. The process document itself is versioned and includes version information for all programs, thus providing a static description of the software environment that should be used to transform input into output files. A process document contains the general layout of a specific type of analysis and has to be created manually by the bioinformatician in charge.

The metadata of an actual execution of the pipeline is captured in the analysis metadata file, i.e., there is a one-to-many relationship between a process document and the corresponding analysis metadata files. Records in an analysis metadata file contain concrete values for all parameters specified as placeholders form in the process document. In conjunction, a process document and a corresponding analysis metadata file represent both the complete set of information necessary to repeat an analysis, and the full documentation of how a specific type of data file is processed in the DEEP consortium to obtain the desired result.



```
<process>
  <name>PROC</name>
  ...
  <desc>...</desc>
  <input>...</input>
  <output>...</output>
  <software>
    <tool>...</tool>
    ...
  </software>
</process>
```

```
[Description]
process        PROCv1
...
[inputs]
input_file_1   run1.in1
...
[outputs]
...
[parameters]
parameter_1    500
...
[metrics]
...
```

```
[Description]
process        PROCv1
...
[inputs]
input_file_1   runN.in1
...
[outputs]
...
[parameters]
parameter_1    1000
...
[metrics]
...
```

Figure 3.1: Documenting pipeline runs with a DEEP process document and analysis metadata files: a DEEP process document (upper left) contains the abstract specification of a standardized computational pipeline. Executing this pipeline with different input data generates distinct output files and metadata records per run (green and orange arrows, top to bottom). The complete metadata record of a run includes the information of input and reference file names (top and right) and is stored in an analysis metadata file (bottom, large icons) .

3.3.2 Concept digest: file tracking database

Every pair of process document and analysis metadata file can be seen as an isolated metadata record that ties the given information to the point in time when the record was created, i.e., when the analysis metadata file was created after a successful run of the pipeline. To ensure a consistent state of this metadata record, it is necessary to link this record to the actual data files in a dynamic manner. We therefore developed a DAC-internal Oracle® database (in the following: file tracking database) that registers DEEP analysis runs and monitors the associated files for changes based on timestamps and data checksums. As long as all data and metadata files do not change, we know that the information captured is still correct at any point in time after the initial registration of the metadata record.

3.3.3 Data exchange between DAC and DCC

As stated in the beginning of this section, the proof of concept status of the DEEP metadata model prohibited a deployment beyond the DAC infrastructure. As a consequence, no error checking system was put into place to avoid the automatic transfer of potentially large amounts of erroneous data from the DCC to the DAC. The download of alignment data from the DCC to the DAC for standardized analysis, and the re-upload of the results from the DAC to the DCC were thus critical points with the potential of introducing errors into the local software environment at the receiving end. Hence, for practical reasons, the consistency of the computational metadata records was only monitored between the time of arrival of new data at the DAC up to the time of the upload of analysis results to the DCC.

3.4 Implementation in DEEP

The specification of the DEEP computational process documents and analysis metadata files was published in Ebert et al. [64]. The description given below includes more technical details compared to the published specification. Furthermore, the description of the file tracking database that establishes the dynamic link between a metadata record and the files on the file system is largely absent from the published manuscript; more details about the file tracking database are also given in the following.

3.4.1 Process and analysis metadata

A process document standardizes one type of computational analysis and contains eight sections (see Table 3.1), which are specified in form and content in an XML Schema Definition (XSD) document. Relying on the well-established XSD/XML formats has the advantage that suitable libraries are available for many programming languages popular in bioinformatics, e.g., R and Python. A crucial point in the decision for the XSD/XML formats was the desired functionality that a DEEP process file can be computationally validated to be compliant with the specification as set in the XSD document. A prototype validation script was published together with the specification in Ebert et al. [64]; for an example usecase, see Listing 3.1. Additionally, the small file size of XML-based process documents — they are just structured text files — facilitates sharing process documents, e.g., by sending them via e-mail to collaboration partners. To improve (human) readability of DEEP process documents, a link to a Cascading Style Sheets (CSS) file is included in each process file. When viewing the process document in a web browser, the CSS document is automatically loaded and the DEEP process document is rendered into a human readable form (Figure 3.2); no special software besides any common web browser and an active internet connection is required.

The sections in the process document are organized into three different categories: metadata about the process file itself (name, version, author, and description; Figure 3.2), a listing of data files (inputs, references, and outputs; Figures 3.3 and 3.4) and a listing of step-by-step instructions detailing the individual computational steps of the process (software; Figure 3.5). Listing reference data as an item separate from input files reflects what we have identified to be a common way of thinking among scientists, i.e., reference data are not subject to data transformations as part of the computational analysis and are thus semantically different from input data. All file listings use generic identifiers (placeholders) to fulfill the requirement of having a single process XML document for all computational analyses of a certain type with run-specific details limited to the analysis metadata file. These placeholders are also used as in the respective command lines in the software section of the process. The computational tools listed in the software section have to be specified with version information to enable other researchers to setup an identical software environment on their infrastructure. Concerning custom scripts, this version information may be stated in form of a repository revision number or commit tag. Beside generic file identifiers, each command line may contain any number of additional placeholders pertaining to parameters whose values depend on the input data. Due to the necessity of using generic file identifiers, the "loop" field explicitly states if all input files are processed together or not, i.e., whether or not the command can be executed in parallel for a set of input files (see Figure 3.5).

The complementary analysis metadata file is automatically produced at the end of a successful analysis run and contains at least the sections description, inputs, references, outputs, and parameters. These sections consist of textual "key-value" mappings that do not require dedicated programming libraries to be created. This lenient format specification is reasonable because all computational researchers in the consortium have to be able to produce analysis metadata files using their preferred programming language in their own software environment. Despite this flexibility, the strict specification of a process document together with the possibility of validating process and analysis metadata together can be used to avoid the collection of unstructured metadata for an analysis run.

Table 3.1: The structure of a DEEP process document. The example name "CHP" is the 3-letter identifier for the DEEP ChIP-seq pipeline.

Section	Content
name	Name of process, e.g., CHP
version	Version of process, e.g., 4
author	Name and e-mail address of process author(s)
description	Free-text description of the purpose of the process
inputs	Listing of all required input files
references	Listing of all required reference files, e.g., genome assembly
outputs	Listing of all generated output files
software	Listing of all analysis steps with details about software

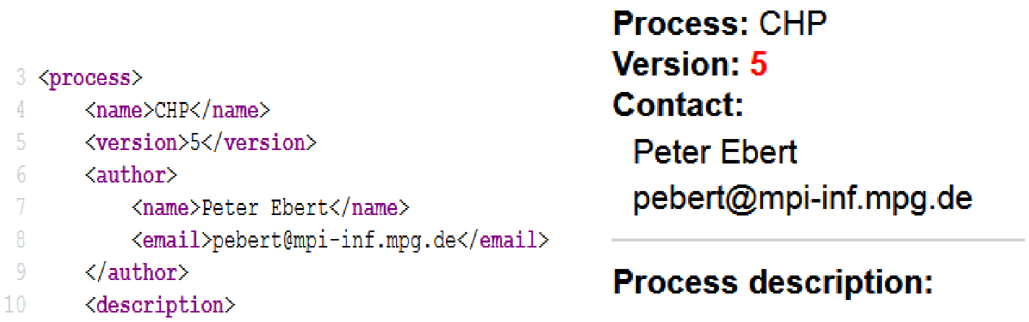
Listing 3.1: Example run of the prototype script validating the process document CHPv5 and the analysis metadata files for several analysis runs. The validation fails for the analysis metadata file that was generated with the previous version of the same process (CHPv4).

```
python3 mdvalid.py \\
  --schema deep_process_schema.xsd \\
  --process CHPv5.xml \\
  --analysis \\
    43_Hm03_BlMa_TO_Hist_F_1.CHPv5.20170916.hg38.amd.tsv \\
    44_Mm03_WEAd_C2_Hist_F_1.CHPv4.20150619.m38.amd.tsv \\
    51_Hf01_BlCM_Ct_Hist_B_1.CHPv5.20170916.hg38.amd.tsv

[INFO] Checking:
 [XML] CHPv5.xml
 [AMD] 43_Hm03_BlMa_TO_Hist_F_1.CHPv5.20170916.hg38.amd.tsv
[INFO] OK:
 [XML] CHPv5.xml
 [AMD] 43_Hm03_BlMa_TO_Hist_F_1.CHPv5.20170916.hg38.amd.tsv
=====
[INFO] Checking:
 [XML] CHPv5.xml
 [AMD] 44_Mm03_WEAd_C2_Hist_F_1.CHPv4.20150619.m38.amd.tsv
[ERROR] Cannot find process name CHPv5
        in analysis metadata section ”description”
[ERROR] AMD file
       44_Mm03_WEAd_C2_Hist_F_1.CHPv4.20150619.m38.amd.tsv
       does not validate against process XML
[INFO] FAIL:
 [XML] CHPv5.xml
 [AMD] 44_Mm03_WEAd_C2_Hist_F_1.CHPv4.20150619.m38.amd.tsv
=====
[INFO] Checking:
 [XML] CHPv5.xml
 [AMD] 51_Hf01_BlCM_Ct_Hist_B_1.CHPv5.20170916.hg38.amd.tsv
[INFO] OK:
 [XML] CHPv5.xml
 [AMD] 51_Hf01_BlCM_Ct_Hist_B_1.CHPv5.20170916.hg38.amd.tsv
```

3.4.2 File tracking database

The file tracking database was designed and implemented in winter 2013/2014 and went online in spring 2014. The database is running on a DAC-internal Oracle® server and all server-side functionality has been implemented in PL/SQL. Programmatic access to the database is realized



```
<process>
    <name>CHP</name>
    <version>5</version>
    <author>
        <name>Peter Ebert</name>
        <email>pebert@mpi-inf.mpg.de</email>
    </author>
    <description>
```

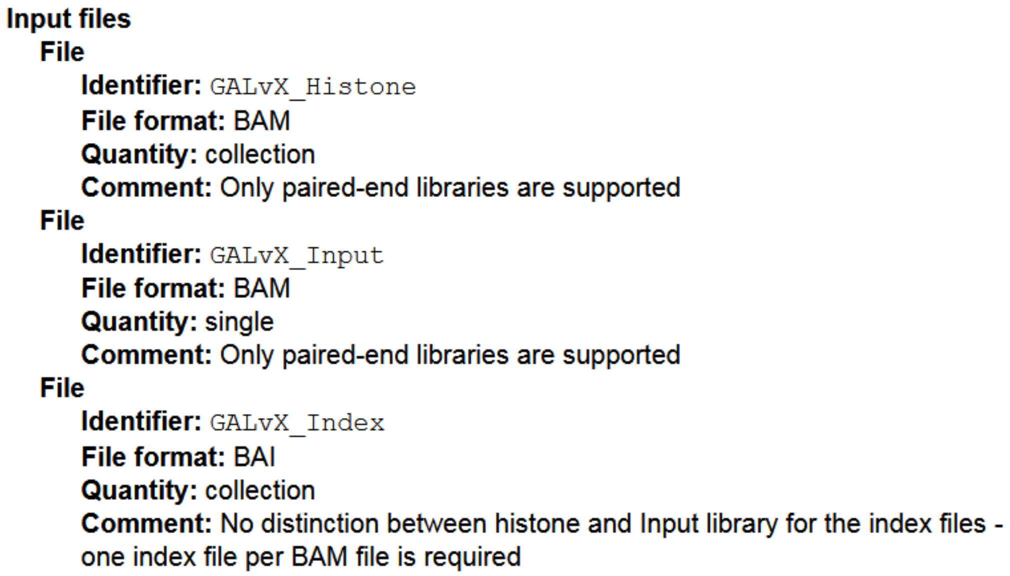
(a) Textual XML view

Process: CHP
Version: 5
Contact:
Peter Ebert
pebert@mpi-inf.mpg.de

Process description:

(b) CSS–rendered view

Figure 3.2: Text and CSS–rendered version of a DEEP process document: excerpt of a DEEP process document in its textual XML form (a) and in its CSS–rendered form (b). The XML document can be opened and read in any text editor, whereas the more human–friendly version requires any common web browser and an active internet connection (rendered here with Mozilla Firefox 61.0.1).

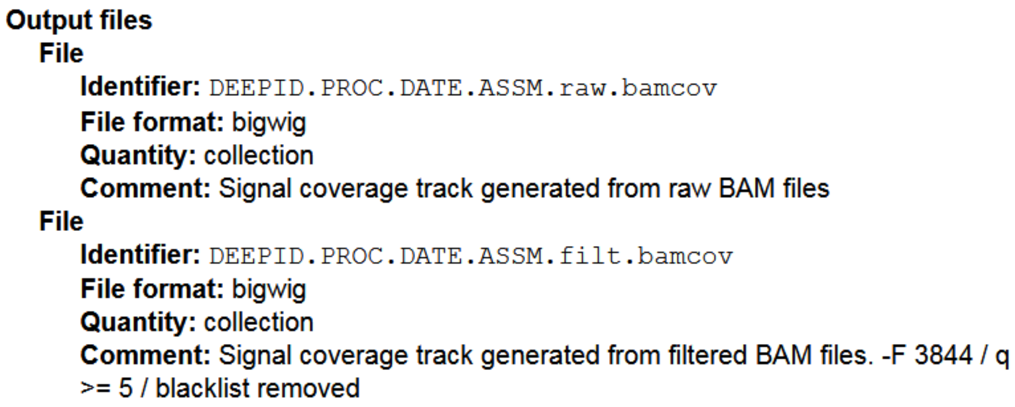


Input files
File
Identifier: GALvX_Histone
File format: BAM
Quantity: collection
Comment: Only paired-end libraries are supported
File
Identifier: GALvX_Input
File format: BAM
Quantity: single
Comment: Only paired-end libraries are supported
File
Identifier: GALvX_Index
File format: BAI
Quantity: collection
Comment: No distinction between histone and Input library for the index files - one index file per BAM file is required

Figure 3.3: Input file listing of a DEEP process document: excerpt of the input file listing in the DEEP process CHPv5. Throughout the process document, files are referenced using a generic identifier. The quantity value “collection” indicates that the process has been designed to handle several files of this type. In this example, the input file identifiers contain a reference to another DEEP process (short read mapping for genomic libraries, GAL), indicating that process CHPv5 is designed for output files of the GAL process.

either directly via a server-side SOAP[17] service, via a custom Python client interfacing with said SOAP service or via a custom Extensible Markup Language Remote Procedure Call (XML-RPC) server that mediates between XML-RPC clients and Oracle’s® SOAP service.

[17] SOAP used to be an acronym for “Simple Object Access Protocol”, but current specifications of SOAP allow more than just object access, hence, SOAP is just a name and no longer an acronym/abbreviation.



Output files

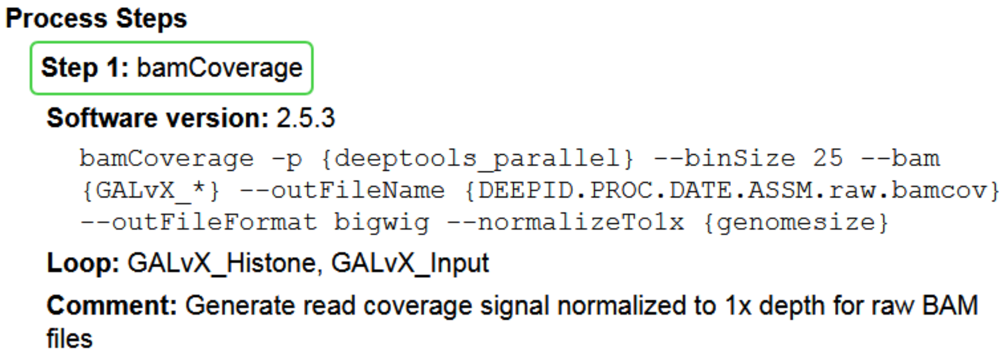
File

Identifier: `DEEPID.PROC.DATE.ASSM.raw.bamcov`
File format: bigwig
Quantity: collection
Comment: Signal coverage track generated from raw BAM files

File

Identifier: `DEEPID.PROC.DATE.ASSM.filt.bamcov`
File format: bigwig
Quantity: collection
Comment: Signal coverage track generated from filtered BAM files. -F 3844 / q >= 5 / blacklist removed

Figure 3.4: Output file listing of a DEEP process document: excerpt of the output file listing in the DEEP process CHPv5. This example illustrates how identifiers for output files should be derived: `DEEPID.PROC.DATE.ASSM.raw.bamcov` is the concatenation of the DEEP sample identifier, the process identifier (here: CHPv5), the date and genome assembly (ASSM) plus a meaningful suffix (here: raw.bamcov) that makes the resulting file name unique.



Process Steps

Step 1: bamCoverage

Software version: 2.5.3

```
bamCoverage -p {deeptools_parallel} --binSize 25 --bam
{GALvX_*} --outFileName {DEEPID.PROC.DATE.ASSM.raw.bamcov}
--outFileFormat bigwig --normalizeTo1x {genomesize}
```

Loop: GALvX_Histone, GALvX_Input

Comment: Generate read coverage signal normalized to 1x depth for raw BAM files

Figure 3.5: Single step of a DEEP process command line listing: excerpt of the software listing in the DEEP process CHPv5. This example illustrates the use of file identifiers as part of the command line specification. The wildcard symbol “*” is used in its common meaning of matching all identifiers with the prefix `GALvX_`, i.e., the histone BAM files `GALvX_Histone` and the Input control `GALvX_Input` are all matched by this expression (Figure 3.3). Parameters in curly braces (`deeptools_parallel` and `genomesize`) may change with each execution of this process and their concrete value has to be specified in the respective analysis metadata file to produce a valid metadata record.

The core components of the database are outlined in Figure 3.6. The database has two main tasks: first, linking files that belong to the same analysis and, second, detecting and reporting file changes after said link has been established. For the first task, the user creates a new record in the database representing an analysis run (table `DEEP.ANALYSISRUN`). This new analysis run is linked to the respective process (table `DEEP.PROCESS`) provided that it already exists. Otherwise, the user is required to create the new process entry. This strict requirement exists because a process serves mainly a documentary purpose in DEEP, and this enforced sequence of steps guarantees that no undocumented analysis can be stored in the database. In a second step, the user supplies lists of input and output file paths that are linked to the metadata record of the analysis run. Here, reference files are subsumed under input files since both file types have to exist before the analysis starts and are, from the database point of view, conceptually identical. The necessary input information to complete these two or three steps has to be supplied manually (see Listing 3.2), but it is of course possible to automate run registration and to add this as the final task of an analysis pipeline; in other words, the system is designed such that human interaction can be reduced to a minimum if desired. After successful analysis run registration, no more manual interaction with the database is required to monitor the coherence of the metadata records.

It is important to realize that this coherence cannot be guaranteed by saving the database entries alone, e.g., in the form of backups. Metadata coherence has to be actively and continuously monitored to safeguard against data corruption. To this end, a daily synchronization between the file tracking database and the DAC file systems scans for changed file modification times. If the timestamp differs between database and file system, the MD5 checksum stored in the database is compared to the MD5 checksum calculated from the file to reliably[18] detect changes of the file's content. Because there are plausible scenarios that do not affect the overall coherence of the analysis metadata records, but yet may result in a different file checksum, e.g., correcting a typographical error, there is currently no automated resolution for confirmed file changes; the emphasis of this setup is that no modification should be unnoticed. Keeping file metadata in a separate location has the additional benefit that, for the duration of the project, recovery after data corruption on the file system level is simplified[19]. Since the database stores filenames and file paths, restoring this information based on a matching checksum is straightforward and faster than complete recovery from — potentially outdated — tape backups. A long-term benefit is that the metadata of a file are kept beyond the file's lifetime on the file system, i.e., if the file is eventually deleted, the database record is set to inactive but otherwise left intact. In case of data inconsistencies detected at any later point in time, this eases the search for the error source as it permits to narrow down the time window when the corruption happened.

The implementation and active use of the file tracking database resulted in other advantageous applications that extended the central tasks described above. The most important ones from the consortium perspective are the automated progress reporting and e-mail notification system, and the possibility to download quality control plots and to annotate samples with quality labels[20]. These services can be regarded as evidence that a system primarily intended to collect and to monitor metadata records can serve various useful purposes in a consortium.

[18] While MD5 is known to be vulnerable and considered insecure for cryptographic uses (see, e.g., the blog post schneier.com/blog/archives/2005/06/more_md5_collis.html by Bruce Schneier), its use to detect data corruption is generally uncontested. At the time of writing, no MD5 collision detected by the file tracking database was due to MD5's known collision vulnerabilities.

[19] Personal communication with IST: despite all redundancies on the hardware and on the software level, there is a non-zero probability that a chain of severe failures results in lost file system metadata, i.e., files still exist, but only as blocks of data with no folder hierarchy, filenames or attributes.

[20] See deep.mpi-inf.mpg.de/status (DAC status page) sections Reporting and Service

Listing 3.2: Example code of using the Python3 SOAP client to store a DEEP process and an associated analysis run of a computational pipeline in the file tracking database. Mandatory parameters are printed all upper case. The complete output of each client call is omitted for visual clarity.

```
$ python3

>>> import deeptrackdb as dtd

# initialize client – pre–configured
# for DAC infrastructure

>>> client = dtd.DEEPTrackDB()

# store new process in database

>>> client.store_process(PROCESS_NAME,
                         PROCESS_FILE,
                         datafreeze=None,
                         freezedate=None)

# create new analysis run
# for existing process
# –> returns DB_RUN_ID

>>> client.store_analysis_run(USER_RUN_ID,
                              PROCESS_NAME,
                              USERNAME,
                              machine=None,
                              cpus=None,
                              runtime_hrs=None,
                              memory_gb=None)

# add files to analysis run

>>> client.store_full_analysis(DB_RUN_ID,
                               INPUT_FILES,
                               OUTPUT_FILES,
                               check=True)
```

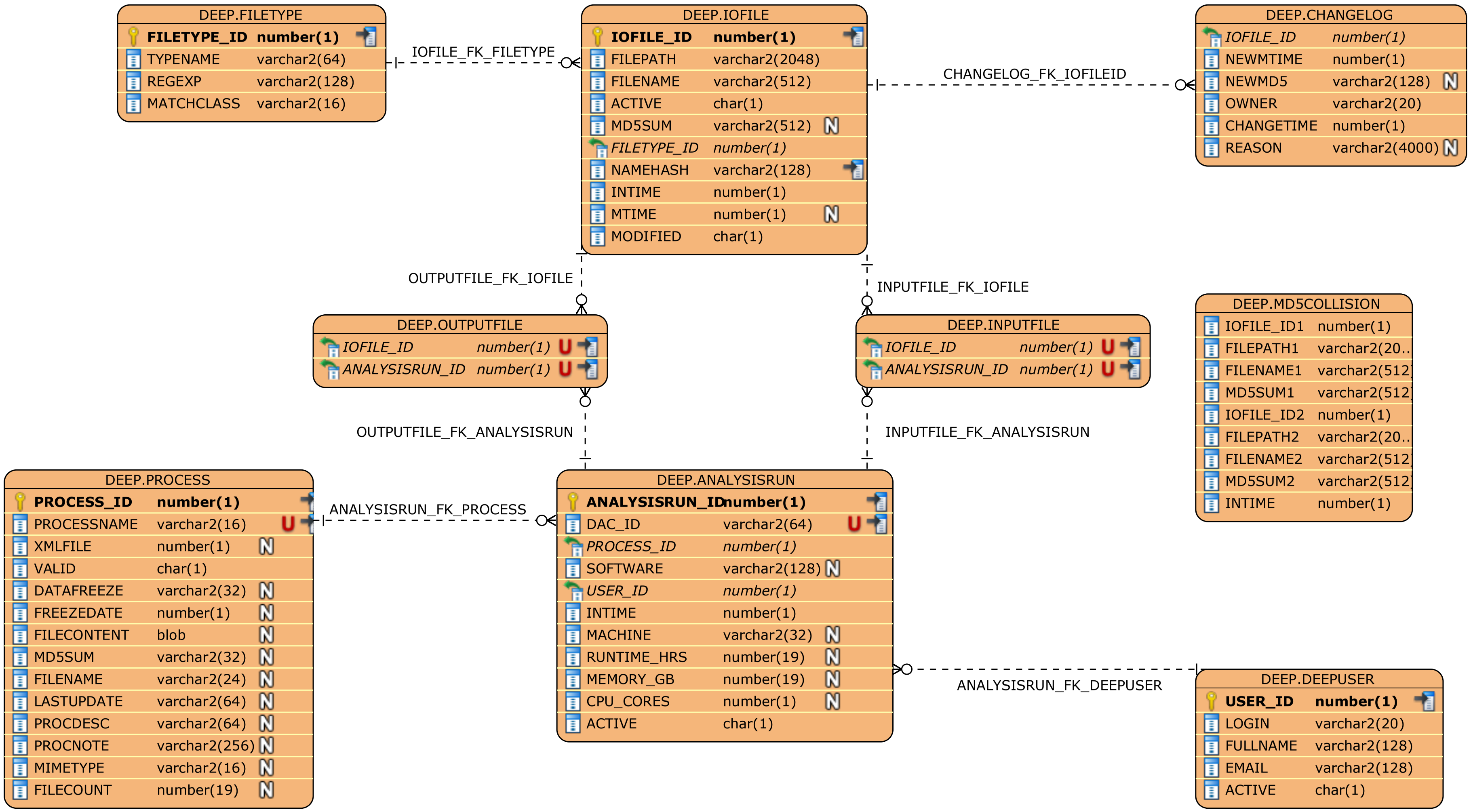


Figure 3.6: Entity-relationship diagram of the core components of the DEEP file-tracking database: tables containing secondary data related to, e.g., progress reporting or the DAC's status page are not shown for visual clarity. Note that foreign key columns (green arrows) are not pointing to the respective key column for layout reasons.

3.5 Evaluation of the Computational Metadata Model

The evaluation of the DEEP metadata concept together with the file tracking database focuses on four major aspects: (i) comprehensiveness of the recorded information, (ii) error detection, (iii) future uses of the system and (iv) discovered shortcomings.

3.5.1 Comprehensiveness of the metadata records

The basic documentation of all DEEP pipelines in the form of process specifications has been continuously updated throughout the lifetime of DEEP. Processes describing pipelines that are not executed by the DAC, e.g., short read alignment or RNA quantification, may not necessarily be in line with current software setups at the respective analysis centers, as this cannot be ensured by the DAC. For all assays routinely analyzed at the DAC[21], the process documents are updated together with the computational pipelines to ensure consistency. At the time of writing, the file tracking database contained metadata records for more than 280,000 individual files and more than 1,000 analysis runs that were registered with an analysis metadata file. For certain types of automatically registered analysis runs, e.g., data down- and uploads between DAC and DCC, no analysis metadata file is created. For these cases, the registration of the analysis run in the file tracking database serves only the purpose of marking the time point when files enter or leave the DAC compute infrastructure. In total, the database records document computational analyses for approximately 250 biological samples[22]. Because the file tracking database did not experience any data loss, the total of all entries can be considered complete from the DAC perspective, and under the constraint that only standardized pipelines are taken into account. The proper documentation of any project-specific analysis of DEEP data is at the discretion of the responsible bioinformatician. However, the database also tracks files not linked to standardized analysis pipelines, so data consistency can be monitored also for data files resulting from custom analyses.

When considering also opportunistic uses of the file tracking database such as the centralized collection of quality labels for individual experiments, it becomes apparent that automation seems key to achieve completeness: the number of manually assigned quality labels is below what would be required to provide a comprehensive picture of the sample quality in DEEP. This gap between mostly automatic data recording and manual effort seems not surprising and highlights the need for research environments that record metadata autonomously and then summarize it in a human-accessible way. Notwithstanding the incomplete sample quality annotation, it seems justified to conclude that the DEEP computational metadata model in conjunction with the file tracking database represents a comprehensive documentation resource that has proven useful in detecting errors and in supporting the global IHEC initiative (see following sections). As such, this project has been a valuable contribution to the overall scientific quality of the DEEP project.

[21] WGBS, RRBS, NOMe, DNase and histone ChIP

[22] This number includes all official DEEP samples plus cell lines and external data that were processed using the DEEP pipelines upon request.

3.5.2 Error detection and consistency checking

Checking the adherence of computational analysis metadata files to the respective process document can be realized using the validation script introduced above (see Listing 3.1). Apart from that, bioinformaticians can rely on the file tracking database for continuous monitoring of the recorded metadata. Since the file tracking database can detect potential errors or data corruptions by itself, an e-mail notification system was implemented that attracts human attention to the file tracking database only when needed. Several examples of these notifications are listed in Table 3.2 with detailed explanations given in Listing 3.5.2.

1. The manual recovery attempt after a failed pipeline run resulted in a misconfiguration specifying the wrong species reference data. The listed binary alignment map (BAM) files have identical checksums despite belonging to human (file 1) and mouse (file 2).
2. File 1 is a pipeline error log, file 2 is a DEEP process document. The error was caused by registering a mock file as DEEP process document in the file tracking database. Both the error log file and the mock process document were empty, i.e., the MD5 checksum is that of an empty file.
3. An error in sample naming and subsequent sorting into wrong folders resulted in data duplication at the DCC. The error was automatically synchronized to the DAC file system.
4. Inconsistent file naming (comma-separated value (CSV) versus tab–separated value (TSV) filename extension) resulted in data duplication at the DCC. The error was automatically synchronized to the DAC file system.
5. An error in sample naming and subsequent sorting into wrong folders resulted in data duplication at the DCC. The error was automatically synchronized to the DAC file system.
6. Both experiment metadata files have the same content, but file 1 belongs to a male subject (“Hm”) and file 2 belongs to a female subject (“Hf”). It should be noted that both filenames are syntactically correct, i.e., there is no direct way to spot this as an error for a human investigator.
7. Despite a difference in the reference assembly used (human GRCh38 and GRCh37), both result files have identical contents. However, this is an example of a “false alarm”: the data files correspond to a microRNA experiment, and microRNAs are not directly mapped against a reference genome assembly, but against a library of potential targets. Hence, a change in the reference assembly may not affect the output files as long as these do not contain actual sequence locations.
8. The contents of the analysis metadata files are identical despite belonging to a steatotic patient (file 1) and a normal control (file 2), respectively. Since analysis metadata files also contain information on the input data files, their contents must not be identical.
9. A case similar to item 6 above, but the two files listed are quality control metadata files produced at the end of a short-read mapping pipeline. That indicates that, due to a wrong sample name, a “new” sample was introduced into the DEEP environment and processed using the default pipelines (again, both sample names are syntactically correct). The corresponding BAM files have likewise been flagged by the file tracking database (not included in Table 3.2).

Table 3.2: Examples of errors detected by the file tracking database that trigger an automatic notification sent to the database maintainer. File paths were reduced to the relevant parts to improve visual clarity. A full description of the notification is given in Listing 3.5.2

No.	Date	MD5 checksum	File 1	File 2
1	2014-06	7087AA76E47EA1194E875B6606990AC7	41_Hf01_LiHe_Ct1_WGBS_S.MCSv0.20140122.bam	44_Mm01_WEAd_C21_WGBS_E.MCSv0.20140204.bam
2	2014-07	D41D8CD98F00B204E9800998ECF8427E	/log/200_mergeOutput.txt	/DEEP/processes/MCSv2.xml
3	2014-11	C91D2F1918AAC5EAA1E14E7FE694C7DC	/43_Hm05_BlMo_Ct_SNRNA_M/[...]	/43_Hm05_BlMo_Ct_snRNA_M/[...]
4	2014-11	ECE7086F1601D4100B87279A630D83A8	41_Hf02_LiHe_Ct2_WGBS_S_emd.csv	41_Hf02_LiHe_Ct2_WGBS_S_emd.tsv
5	2014-11	44D41164627F4B8CE9D499CE9EBBE991	/43_Hm01_BlMo_Ct_SNRNA_M/[...]	/43_Hm01_BlMo_Ct1_SNRNA_M/[...]
6	2015-12	B3CE2C00FF639EB6EF437B859FD47E86	52_Hm3b_CoMu_UC_NOMe_S_1_emd.tsv	52_Hf3b_CoMu_UC_NOMe_S_1_emd.tsv
7	2015-12	287FE70BD3F900C2A352A796223B7A06	/GRCh38/[...]/41_Hf11_LiHe_St_snRNA_K_1	/GRCh37/[...]/41_Hf11_LiHe_St_snRNA_K_1
8	2015-12	9947CDA1F03A68EAC41E291358136552	41_Hf05_LiHe_St_snRNA_K_1.SALv2.20150922.amd.tsv	41_Hf03_LiHe_Ct_snRNA_K_1.SALv2.20150922.amd.tsv
9	2015-12	5AAC882F877AD0088DE854E1CCCEF0A2	51_Hf07_BlTM4_Ct_WGBS_S_1.QcSummary.20151116.txt	51_Hf07_BmTM4_Ct_WGBS_S_1.QcSummary.20151116.txt
10	2016-07	—	41_Mm11_LiNP_OS_smd.txt	41_Mm11_LiNP_OS_tRNA_K_1_emd.tsv

10. An example of a file change notification after an update of several sample and experiment metadata files at the DCC. The changed file modification time triggered the comparison of the MD5 checksum stored in the file tracking database to the newly calculated MD5 checksum of the file; since the two checksums differed, the notification was sent to indicate a potential data corruption.

An instructive example of a case where the existing software solutions could not have detected a problem occurred in May 2015 as follows: after a routine update of the DEEP ChIP-seq pipeline for basic signal quantification and peak calling, one of the quality control plots depicting a correlation heatmap between all histone marks showed a substantial deviation from the previous version. Figure 3.7a shows the original heatmap created in May 2014, and Figure 3.7b shows the updated version created in May 2015. The two heatmaps clearly show differences suggesting that the original results cannot be reproduced (or rather replicated) with the updated computational pipeline. Due to the extensive metadata records at the DAC, it was possible to conclude beyond reasonable doubt that this anomaly did not result from an error in the DAC processing pipelines. The source of the problem was traced back to a change in deepTools [199], where a previously implemented stringent strategy of outlier removal was changed without documenting this in the release change logs (see Figure A.1). It is a matter of personal experience whether one considers a minor negligence like missing documentation of changed program behavior a common situation in computational research or not. Irrespective of personal experiences, though, without the metadata records kept at the DAC, it would not have been possible to quickly isolate the source of the problem with reasonable certainty, which would have put all previous results into question.

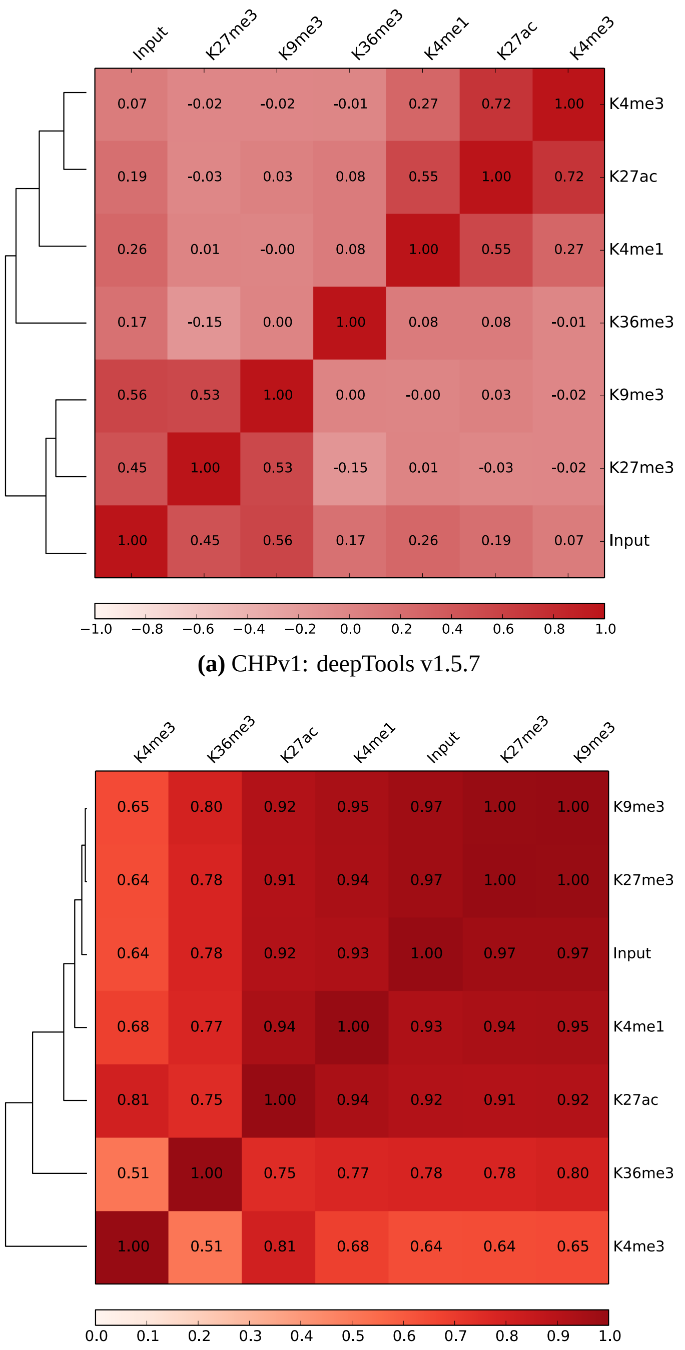
3.5.3 Post-DEEP use of the metadata records

The funding period of the DEEP consortium ended in October 2017 and the file tracking database can be archived at the DCC in Heidelberg as soon as all remaining DEEP analyses have been completed. Since all records can be dumped as SQL or textual tables, it will still be possible to retrieve all information without the need for the complete programmatic environment described in section 3.4.2. The documentation resource created at doi.org/10.17617/1.2W is linked via the IHEC data portal [36][23] and provides open access to the complete description of the computational pipelines developed and applied in DEEP. It follows that full documentation and metadata records will be available to the scientific community beyond the lifetime of the DEEP project.

3.5.4 Identified shortcomings in design and implementation

The routine use of the DEEP metadata model revealed several weaknesses of the setup. The requirement of straightforward, iterative updates of the process documents and of the associated pipelines turned out not to be in line with realistic needs. Empirically, process documents and the respective computational pipelines were usually only updated when reference data changed, e.g., after switching to a new version of a genome reference assembly. This raised the question

[23] Documentation linked under “Methods” for the DEEP consortium.



(a) CHPv1: deepTools v1.5.7

(b) CHPv3: deepTools v1.5.9.1

Figure 3.7: CHPv1 and CHPv3 quality control heatmaps: (a) histone mark Pearson correlation heatmap for mouse sample 44_Mm01_WEAd_C2 created with deepTools v1.5.7 as part of DEEP process CHPv1 in May 2014. (b) the same data were analyzed with deepTools 1.5.9.1 as part of the updated CHPv3 process in May 2015. Rows and columns of both heatmaps are labeled with the respective histone mark name, "Input" refers to the Input control dataset.

whether the flexibility of specifying reference data was actually necessary; tying a fixed set of reference datasets (per species) to one process version may be sufficient to fulfill the needs of a large-scale consortium with a predefined research scope. Of course, a solution that also ties specific reference data to process versions requires consortium-wide agreements which references to use for every single reference annotation file.

A stricter handling of user-submitted data was also discussed as a potential enhancement of the file tracking database with the objective of improving metadata consistency. For example, the file tracking database could have canceled the registration of analysis runs if process document and analysis metadata failed to validate. The downside of such restrictive approaches is that usability usually suffers, which in turn can be expected to lower user compliance. However, since the metadata model was developed as part of DEEP, its use in daily operations can be considered a field test to detect deficits; to that end, a lenient system potentially offers more insights into what features users actually expect, want or use.

One of the design principles behind the DEEP metadata model is the reliance on rather simple data formats to facilitate data exchange between all collaborating institutes. Consequently, DEEP process documents are text files and not, e.g., executable pipeline scripts. Nevertheless, it would have been possible to extend the tooling support for the computational metadata beyond the validation script. For example, it would likely be feasible to implement a tool that extracts the command lines from the process documents to provide bioinformaticians with a rudimentary (Linux) shell script. Although such auto-generated basic shell scripts are presumably not out-of-the-box portable across the infrastructure of all partner institutes, they could have been useful for small-scale testing of updated (parts of) pipelines or for running a DEEP pipeline on selected non-public/non-consortium datasets.

The last potential deficiency concerns the better integration of additional metadata related to individual samples and experiments. The file tracking database extracted information about DEEP samples exclusively from the filenames and file system folder structures. An explicit inclusion of sample and experiment metadata in the file tracking database could have helped to identify missing data files and missing metadata records (see Section 3.6.3 below). It could be argued that this would turn the file tracking into a sample tracking database, which is out of scope for the DEEP computational metadata model. In particular, the computational validation of sample and experiment metadata can be cumbersome — if not impossible — without strict specifications of controlled vocabularies or biological ontologies. It follows that, as a prerequisite to the explicit inclusion of additional metadata in the file tracking database, collection and validation strategies have to be defined in more systematic ways. Given that groundwork, augmenting the file tracking database with sample and experiment metadata would be feasible and, as will be argued in Section 3.6, would also represent an important building block for reproducible research[24].

[24] The IHEC Data Ecosystem working group is developing strategies to validate experiment and sample metadata using JSON and XSD/XML based templates similar to the DEEP computational metadata model; see github.com/IHEC/ihec-ecosystems/tree/master/version_metadata/examples

3.6 Discussion and Perspectives

The final section condenses the central aspects of this chapter into a tentative outlook on future strategies for reproducible (computational) research. Some points made in this discussion may be perceived as optimistic, but the practical suggestions that could be — and partially were — implemented in a research consortium such as DEEP only require comparatively little changes to established operating procedures. It is the premise of this discussion that gradual improvements in reproducibility should be preferred over absolute — but hard to achieve — "quantum leap" solutions.

3.6.1 Is the concept of reproducibility outdated?

The reproducibility crisis and the apparent lack of a self-evident way out of it motivate the question if scientific reproducibility is still an appropriate quality measure for data-driven, high-throughput biomedical research in the 21st century. The consensus in the current literature seems to be that only reproducible research is reliable research [4, 19, 47, 61, 103, 108, 137, 228]. The generally accepted answer to the question why that is the case is that testing a new hypothesis requires a framework of previously established theories that are assumed to hold (until proven otherwise, see Section 3.1); science can only advance in a systematic manner under that condition.

However, accepting the above answer to the question why reproducibility is still fundamental to (modern) scientific progress does not provide an immediate answer to the question what that entails; two parts seem to be relevant here: (i) what studies need to be independently tested and (ii) to what extent must the results be concordant to call the original study reproducible? Concerning the first part of the question, an observable pattern is that controversial or groundbreaking studies are often reproduced by independent groups out of scientific interest; see, e.g., the case of Obokata et al. [175], where the results could not be reproduced and the publication was retracted[25], or the so-called "Schön scandal"[26]. It thus seems reasonable to assume that the majority of published research is left untested, and the published claims or findings are, from a theoretical point of view, in an unclear state of reliability. A conceivable reason behind this is a lack of broad scientific interest, which justifies not to raise the necessary resources to reproduce any published body of work.

The second part of the above question concerning concordance of reproduced and original results can be assumed to be (even) more elusive, because the required (minimal) agreement between an original study and its successful reproduction is presumably context-dependent [244][27].

In summary, it seems difficult to find a precise answer to the question what the importance of reproducibility — if anything at all — implies for modern high-throughput research. Reality offers at least a practical answer, because science is advancing despite a lack of independent testing for most studies; some even argue that most published research findings are false [107]. However, it

[25] Nature retraction notice: doi.org/10.1038/nature13598

[26] Nature News: doi.org/10.1038/news020923-9

[27] A failed or succeeded attempt to reproduce a study is presumably often a case of "I know it when I see it".

may be that speed and resource-efficiency of the current scientific progress are perceived as less than adequate [152]. In conclusion, reproducibility as a condition for scientific progress seems not to have lost its importance for modern research, but the concept of scientific reproducibility may be in need of a "theory of sufficiency" [80] to keep pace with changing trends in science (e.g., from hypothesis-driven to data-driven science). A theory of sufficiency[28] for scientific reproducibility should define sufficient conditions under which a scientific field can advance without aiming for potentially wanted but hard to achieve ideals such as reproducibility rates of 100%. Defining a theory of sufficiency for scientific reproducibility is of course not the objective of this thesis. Hence, the following paragraphs focus on more practically oriented suggestions for modernizing some aspects of scientific work with a potentially positive effect on reproducibility.

3.6.2 Sufficiency in reproducibility as a community-driven project

The driving force behind any change in the contemporary implementation of the scientific method has to come from within the scientific community, which includes all involved parties, from journal editors and reviewers to principal investigators, staff scientists, and data curators. As pointed out in Section 3.1.3, formulating demands on how to improve reproducibility that address the scientific community as a whole seems not to be the right way of solving the reproducibility crisis. A reasonable alternative would be to tackle the problem at smaller scale on the level of subcommunities. Subcommunities of comparatively narrow research fields, e.g., ChIP-seq, regulatory network analysis or short read mapping could regularly assemble to host *Quo Vadis* events. During these gatherings, experts and proficient users could jointly identify core issues in their niche: what published work is of general interest and should be reproduced? What open questions should be addressed to move the field forward? What databases or data resources are widely used and need curation, better interfaces or sustained maintenance? What data standards are needed, or need an update, or are obsolete altogether? The answers to these questions could then be summarized in a white paper[29] that serves multiple purposes: first, it highlights problem areas relevant to the community, which may help to acquire funding for low-profile and yet important work such as infrastructure or maintenance projects. Second, it identifies research problems that are of broad interest to the community, which may provide guidance to make better decisions at crucial points during a research project[30]. Third, the research community outlines what published research needs to be independently tested to provide proper footing for future work. The actual reproduction work could then be carried out by volunteering labs that receive earmarked funds for each reproducibility study. Although this community-driven process does not by itself constitute a "theory of sufficiency" for modern reproducibility, it yet accomplishes three things in this regard: first, it locates concrete responsibility and action within a small group of researchers

[28] The cited source by Harry G. Frankfurt deals with sufficiency in economics, which is of course not relevant here. However, using the term and the concept of a "theory of sufficiency" was motivated by this source.

[29] The Computational Pan-Genomics Consortium has published a white paper along those lines [157]. The purpose of this white paper was to provide an overview of this emerging field, and thus the manuscript does not address potential reproducibility issues; this could of course change as soon as the field is more established.

[30] An obvious objection against such an approach would be that it could limit diversity of research or amass to much resources on a narrow scope. Although theoretically this could happen, it should be noted that, by complementation, such a white paper also points into directions of high-risk/high-gain research.

instead of addressing the scientific community as a whole (Section 3.1.3). Second, as an immediate consequence, the scientific community anchors reproducibility studies as an integral part of the scientific work routine, which attributes more value it. Third, instead of considering (the lack of) reproducibility as an absolute problem, an iterative and adaptive strategy sets short-term goals of what is needed for any particular field to advance “in good faith” — the scientific endeavor as a whole would then benefit from these growing fields of “sufficient reproducibility” established at the level of subcommunities.

3.6.3 Reproducibility by design in large-scale research projects

Under the assumption that a notion of sufficiency in reproducibility is a practical necessity given the current pace of research, it seems pertinent to examine potential implications for large-scale research projects such as DEEP. It seems indisputable that the main objective of research consortia would not be altered by a more contemporary view on reproducibility: the expertise of all collaborating groups is combined to tackle difficult or even high-risk research problems that are beyond the reach of any individual group. Hence, the studies published by the research initiative are the central output. As argued in Section 3.6.1, if the published results are groundbreaking or controversial (enough), it can be expected that the results will be independently tested by groups outside of the consortium. For work with lower impact, the realization of *Quo Vadis* events including reproducibility studies by volunteering labs would offer another route to gaining higher confidence in results published by research consortia. However, it does not seem reasonable to assume that the majority of the works published in a consortium context would be independently tested for their reproducibility. It follows that some of the findings reported by a research consortium are of unclear value to the scientific community (Section 3.6.1). This does not by any means diminish the individual achievement; yet it suggests that large research consortia may need to modernize their role as providers of comprehensive datasets to support (sufficient) reproducibility by design. The common practice of depositing raw data at an International Nucleotide Sequence Database Collaboration (INSDC) [211] database may ensure long-term availability of the data, but this is not *per se* a useful resource. Although a certain amount of annotation or metadata is required before a raw dataset can be deposited at an INSDC database, consortium-internal rules on sample naming, recorded sample information or trivialities like units of measurement may complicate the creation of a coherent data resource at an INSDC archive.

A possible improvement over the current situation could be achieved by collecting sample, experiment and computational analysis metadata in a consortium-internal database that allows for controlled access by external experts for “virtual reproducibility audits”. During such an audit, external experts would examine the coherence of the metadata records with the objective of “virtually” reproducing all analysis steps. In other words, the external experts would try to evaluate if the metadata recorded for a specific analysis task are presumably sufficient to independently reproduce the results given suitable primary data. Due to their experience, experts can spot missing but relevant metadata, or identify potential low quality samples, e.g., based on quality control metrics without having to repeat the analysis. These audits could be conducted as part of standard reporting procedures, and potentially more often upon request. Furthermore, they would have several beneficial properties: first, the burden on the infrastructure is low as the data volume is limited — metadata have usually the form of textual information — and the necessary technology, i.e., databases with simple and access-controlled interfaces is well-established. Second, the consortium does not suffer from a competitive disadvantage because no access to the primary data is required. Third, the outside perspective prevents the effect of "organizational blindness" when evaluating comprehensiveness and usefulness of the metadata records. Fourth, potential missing information would be discovered before data submissions to INSDC databases have to be prepared, which may increase the chances of collecting the missing information before data submission. Fifth, in emerging fields with no consensus on the reporting standards, this transparency may help to identify a core set of sample, experiment or computational analysis metadata that need to be recorded irrespective of the study context. Sixth, if the medium- to long-term preservation and accessibility of the consortium metadata records cannot be ensured by any of the collaborating groups, saving the records at a dedicated service such as Zenodo[31] could solve the issue. In summary, planning for "virtual reproducibility audits" when devising resource requirements for large-scale research projects may offer a practical way for "sufficient reproducibility" by design within the sphere of influence of a consortium.

3.6.4 A building block for reproducible research in DEEP

The DEEP metadata model for computational analysis is the realization of one of the central building blocks that are required for preparing a consortium for "virtual reproducibility audits". The proof of concept status of the metadata model, including metadata validation and the file tracking database, prevented a more strictly enforced — instead of just encouraged — use throughout the DEEP consortium. Due to this proof of concept status, the last objective formulated in Section 3.3, i.e., the continuous and centralized collection of all metadata at the DCC in Heidelberg, has not been realized. Nevertheless, the remaining objectives outlined in Section 3.3 have been achieved: the metadata model described in this chapter is a task-oriented approach for recording computational metadata in a large-scale research consortium. This model requires only limited manual work, i.e., specifying process documents, but captures a substantial amount of information when considering the combination of process documents and analysis metadata files. Validating these metadata records (Listing 3.1) and continuously checking their consistency via the file tracking database ensures that this body of information provides a reliable resource for future queries about the DEEP data. Specifically, this setup ensures that hypotheses about the data can be tested with certainty about data processing steps, which can provide important hints for elucidating the nature of spurious observations. The DEEP computational metadata model abstracts implementation-specific intricacies of analysis pipelines into a human- and machine-readable form without compromising on the comprehensiveness of the metadata records. To that end, the model is designed to require only limited resources while overall capturing a sufficient amount of information such that reproducing (computational) results is a practically achievable goal. It is self-evident that the DEEP computational metadata model cannot prevent human error (but it can detect it, see Section 3.5)

[31] Zenodo is a general purpose, "catch-all" repository for scientific data funded by the European Commission and hosted by the CERN.

or negligence, and, consequently, cannot provide guarantees on the reproducibility of a study. It does, however, provide reliable information on the "what and how" for individual results. It is this combination of lowering error rates and providing reliable information that harbors the potential of substantially improving reproducibility.

CHAPTER 4

Fast Detection of Differential Chromatin Domains with SCIDDO

Lead-in This chapter is concerned with the problem of systematically identifying regions of differential chromatin between cellular phenotypes of interest within one species. Section 4.1 provides a short motivation and an overview of existing methods that tackle the same problem. As detailed in Section 4.1, these existing methods have several methodological limitations that motivated developing our tool SCIDDO as a more broadly applicable and more flexible approach toward differential chromatin analysis. Section 4.2 presents the methodological and theoretical foundations of SCIDDO, followed by the results of an exemplary study that illustrate potential use cases for SCIDDO (Section 4.3). The results are discussed in Section 4.4 and some pointers for future work on SCIDDO are included in Chapter 6 of this thesis.

The work presented in this chapter is an extended version of the manuscript Ebert and Schulz [63] (for details and author contributions, see Appendix E.1.2).

4.1 Background

Large epigenome mapping consortia such as DEEP, BLUEPRINT or ENCODE produce an ever-increasing amount of reference epigenomes for a multitude of different cell types. With the ultimate goal of compiling a publicly available catalog of 1,000 reference epigenomes released under the IHEC umbrella, comparative analyses of epigenomes present a formidable challenge for bioinformatics. However, as explained in Chapter 2, the cell-type specific and dynamic nature of the epigenome adds substantial complexity to the problem of characterizing cellular similarities and differences on the epigenetic level. Moreover, limited resources commonly force scientists to investigate only a small number of biological replicates per condition of interest. Despite these challenges, the discoveries in the field of epigenomics have greatly enhanced our understanding of transcriptional regulation, cellular identity and disease development [23, 102, 114, 135, 149].

Following the main topic of this thesis, this chapter focuses on the histone chromatin component of the epigenome. The interpretation of histone mark data is particularly intricate as the interplay between different histone marks results in a combinatorial complexity that is largely absent for other epigenetic modifications such as DNA methylation (see, e.g., bivalent domains described in Chapter 2, Section 2.1.1.4 and Table 2.1). The realization that histone mark combinations can be interpreted as local activity states of the genome, so-called chromatin states, led to the widespread use of probabilistic graphical models for discovering these "hidden states" [70, 72, 105, 154, 230]. Popular tools such as ChromHMM [70] or EpiCSeg [154] have tremendously simplified the analysis of histone data as they summarize the combined effect of histone mark co-occurrences in a manageable number of discrete chromatin states. After functional characterization, the discovered chromatin states are commonly augmented with textual labels to ease interpretation, e.g., identifying regions as active or poised promoters (see Chapter 2, Section 2.1.1.4 and Table 2.1), or distinguishing between weak and strong transcriptional activity. However, in our experience, the generated chromatin state maps are often manually inspected in only a limited number of loci or simply serve as additional genomic annotation data. Given that chromatin state maps provide a neat abstraction of the various histone mark combinations, it stands to reason that a more comprehensive view on them may offer valuable guidance in exploratory studies.

So far, there are only few tools available that use chromatin state maps to identify regions of differential chromatin marking. ChromDet [39] can be applied in a genome-wide manner and uses Multiple Correspondence Analysis (an analog to Principal Component Analysis for categorical data) followed by an iterative clustering approach to identify regions that partition the samples into cell-type or lineage specific groups (so-called chromatin determinant regions). The computational burden of a ChromDet analysis is lowered by various filtering steps to remove, e.g., uninformative or outlier regions, which renders ChromDet analyses prohibitive for small sample numbers[1]. This preprocessing also requires enough insight into the nature of the samples at hand to manually set appropriate filtering thresholds.

Other available tools for the differential analysis of chromatin state maps enable only the analysis of a predefined set of genomic regions. The ChromDiff [264] tool first computes the percent coverage of all chromatin states in each user-specified region of interest, e.g., the bodies of all coding genes. Next, after correcting for batch effects using a regression model, ChromDiff uses the non-parametric Mann-Whitney-Wilcoxon[2] test [155] to identify differential chromatin states between sample groups, e.g., contrasting all male and female samples. Since ChromDiff relies on standard statistical tests for its analysis, sufficient statistical power in terms of number of available samples per group is mandatory to find any significant differences between the groups. The recently published Chromswitch package [111] similarly identifies differential chromatin states only in preselected regions of interest. Chromswitch can only analyze a single chromatin state at a time and uses a binary "presence/absence" encoding to construct feature vectors that are subsequently clustered. The cluster assignments resulting from the hierarchical clustering

[1] Personal communication with the developer in March 2018 via GitHub: the minimum number of samples required for an analysis presumed reliable is three biological replicates for three different cell types. Ideally, the dataset should consist of 5–10 biological replicates for four or more cell types. For datasets with lower numbers of replicates per group, ChromDet may finish the analysis with the status "not enough group diversity" and not return any output.

[2] The test name stated here is as given in the publication by Yen and Kellis [264]. This is the same test as the Mann-Whitney-U test that is used several times in this thesis.

are then scored by their agreement with the known biological labels of the samples and manual thresholding on these scores is required to select the final set of differential chromatin regions.

A common denominator of all surveyed methods is that they consider chromatin state similarity as a binary variable, i.e., any chromatin state is (dis-) similar to any other chromatin state to exactly the same extent. We think that this is an unnecessary simplification that does not fit the commonly encountered interpretation of chromatin states as representing different activity levels of the genome. For example, it seems counterintuitive that chromatin states representing strong and weak enhancer activity should be as dissimilar to each other as they are to a state representing heterochromatin.

In summary, current methods are limited to region-based analysis, focus on individual chromatin states, require a comparatively large number of biological replicates for their statistical analysis, and use a quite basic representation of chromatin state similarity, which hinders general applicability of existing methods.

We devised a new method for the score-based identification of differential chromatin domains (SCIDDO) with the goal of providing a generally applicable tool for the fast identification of differential chromatin marking. One of SCIDDO's main features is its capability of identifying potentially large and heterogeneous regions of differential chromatin marking, which we refer to as differential chromatin domains (DCDs). The statistical evaluation of the identified domains relies on well-established theory borrowed from score-based biological sequence analysis. This enables an interpretable presentation of SCIDDO's results and facilitates downstream analysis. Moreover, SCIDDO enables users to use custom scoring schemes for an analysis, offering a flexible way of defining task-oriented and quantitative notions of differential chromatin. In the following, we present results obtained by analyzing four groups of replicated human samples with SCIDDO. In this analysis, we assessed the robustness of our method by comparing DCDs between individual replicates and characterized the identified domains by overlapping them with differentially expressed genes (DEGs) and various regulatory annotation datasets. We compared SCIDDO to other methods for the differential analysis of histone data and collected evidence that highlights SCIDDO's usefulness in identifying regions of dynamic chromatin changes, e.g., enhancers switching from an "on" to an "off" state between cell types.

4.2 Materials and Methods

4.2.1 Experimental data overview

All analyses were carried out using the official IHEC human hg38/GRCh38 assembly. We selected the following high quality DEEP samples to include both closely related as well as more distantly related cell types in our analysis: two replicates of HepG2 (HG 1 and 2; Table B.4: Online Table S1), two replicates of hepatocytes (He 2 and 3; Table B.4: Online Table S1), three replicates of monocytes (Mo 1, 3, and 5 [254]) and two replicates of macrophages (Ma 3 and 5 [254]). All primary cell types were isolated from healthy, adult donors. For each replicate, we downloaded the DEEP reference alignments for six histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3, H3K9me3) and the corresponding Input[3] control as BAM files (Table B.4: Online Table S1). Additionally, we downloaded DEEP mRNA expression data for all samples as raw read FASTQ files (Table B.4: Online Table S2). The hg38 genome reference was restricted to fully assembled auto- and gonosomes for all data preprocessing steps. The differential analysis with SCIDDO was then limited to autosomes and chromosome X to alleviate any effects arising from the uneven distribution of sexes in our dataset. Annotation data were likewise limited to the same set of chromosomes. The GeneHancer [78] enhancer annotation was licensed for academic use on 2017-05-30. The GeneHancer annotation was reduced to gene-enhancer pairs that could be mapped to gene identifiers in the GENCODE v21 annotation [96].

4.2.2 Generation of chromatin state maps

Following IHEC recommendations, all histone BAM files were filtered using Sambamba v0.6.6 [237] to exclude low quality reads (mapping quality ≥ 5; no duplicated, unmapped or non–primary reads/alignments). These filtered BAM files were used as input to generate chromatin state segmentation maps for all samples. We used a pre-trained ChromHMM model provided by the REMC. This model was trained to segment the genome into 18 chromatin states, hence we refer to this model as CMM18. We decided to use the pre-trained CMM18 model because it has been carefully designed using the large compendium of epigenomes generated by the REMC. We thus assumed that CMM18 robustly captures chromatin states irrespective of the biological source of the samples at hand (we examined this assumption using a newly trained ChromHMM model, see Section 4.2.3). As an additional benefit, the chromatin states of the CMM18 model were functionally characterized and labeled by the REMC to facilitate interpretation of the state segmentation maps (for state descriptions and colors, see Figure B.1 and Table B.1). We executed version 1.12 of ChromHMM with commands “`BinarizeBam -b 200`” and “`MakeSegmentation -b 200`” and otherwise default parameters to create the state segmentation maps.

4.2.3 Applicability of the pre-trained CMM18 model

We used ChromHMM version 1.12 to train a new 18-state model on our dataset with the command “`LearnModel -b 200`” and otherwise default modeling parameters. We then compared the chromatin state emission probabilities of the newly trained model (NEW18) to the state emission probabilities of the pre-trained model (CMM18) provided by the REMC. The task of matching the chromatin states between the NEW18 and the CMM18 model was modeled as a “minimum weight perfect bipartite matching” (in the following: “assignment”) problem and solved using linear programming. This assignment problem can be conceptualized via a bipartite graph consisting of the vertex sets A and B, where each of the 18 nodes in A and B represents one chromatin state of the NEW18 and of the CMM18 model, respectively. The edges E in this bipartite graph are given as $E = A \times B$. For each edge e_{ab} connecting vertex $a \in A$ to vertex $b \in B$, we computed the edge weight w_{ab} as the Kullback-Leibler divergence (KLD) [143] between the state emission

[3] Reminder: the capitalized Input designates the ChIP-seq Input control sample as introduced in Chapter 2, Section 2.2.2.

probability distributions of a and b. The KLD can be interpreted as the increase in entropy if, in our case, the chromatin state predicted by the NEW18 model is approximated by the predefined chromatin state from the CMM18 model. A weight w_{ab} was then computed as

$$w_{ab} := KLD(a, b) = \sum_i a_i \cdot log\left(\frac{a_i}{b_i}\right) \tag{4.1}$$

with i referring to the individual components of the emission probability distribution, i.e., in our case, the observed six histone marks that were used as input data to generate the chromatin state segmentation maps (Section 4.2.2). The assignment problem can now be formulated as the following constrained minimization problem:

$$\begin{aligned}
\text{minimize:} \quad & \sum_{a \in A} \sum_{b \in B} w_{ab} \cdot e_{ab} \\
\text{subject to:} \quad & \\
& \sum_{b \in B} e_{ab} = 1 \text{ for } a \in A \\
& \sum_{a \in A} e_{ab} = 1 \text{ for } b \in B \\
& e_{ab} \geq 0 \text{ for } a \in A \text{ and } b \in B
\end{aligned}$$

Solving this problem amounts to selecting 18 edges that match each chromatin state of the NEW18 model to exactly one state of the CMM18 model such that the sum over all edge weights is minimal and no two edges are incident to the same vertex $b \in B$. The problem was implemented and solved with default configuration in the Python PuLP package[4]. As comparison, we also solved this problem in a greedy way by iteratively selecting the edge e_{ab} with minimum weight w_{ab} if no previously selected edge was already incident to a or b.

4.2.4 Differential gene expression analysis

Gene expression estimates per replicate were computed with Salmon v0.9.1 [180] using the GENCODE v21 [96] annotation for protein coding genes. For each gene in the GENCODE reference, we extracted genomic coordinates for the gene body (5' to 3' end) and for the promoter (-2500 bp to +500 bp around the 5' end) using custom scripts. After expression quantification, we used DESeq2 v1.18.1 [147, 229] to obtain differential expression estimates for all six possible pairs of sample replicate groups in our dataset. We split the DESeq2 results into groups of DEGs and non-differentially expressed genes (stable genes) based on an absolute log2 fold change in expression of at least 2 and a multiple-testing corrected p-value of less than 0.01 (DESeq2 p-value correction method: Benjamini-Hochberg [20]).

[4] PuLP v1.6.8: github.com/coin-or/pulp

4.2.5 Differential histone peak calling

We selected PePr [267] as a current state-of-the-art tool for differential chromatin analysis as reference to compare to. We executed PePr v1.1.18 to perform differential analysis including postprocessing for all six possible pairs of sample replicate groups in our dataset. All available replicates were processed in a single run of PePr for each comparison. PePr was executed with the parameter histone “`peaktype`” set to “`broad`” for the mark H3K36me3, and otherwise default parameters (see Chapter 2, Section 2.4.2 for histone peak types). The resulting histone peak sets were reduced to those peaks with a multiple-testing corrected p-value of less than 0.01 using custom scripts (PePr p-value correction method: Benjamini-Hochberg [20]). In the comparison between SCIDDO and PePr, the method performance in detecting DEGs is evaluated using accuracy and F1 score, the latter being the harmonic mean of precision and recall. The following definitions [183] use the common shorthand notation P to denote all positive samples, N to denote all negative samples, and TP, TN, FP, FN to denote true positives, true negatives, false positives, and false negatives:

$$\begin{aligned} Accuracy =& \frac{TP + TN}{P + N} \\ Precision =& \frac{TP}{TP + FP} \\ Recall =& \frac{TP}{TP + FN} \\ F1 =& 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision} \end{aligned}$$

4.2.6 Chromatin dynamics at EP300 peaks

SCIDDO implements an optional postprocessing step for restricting the set of DCDs to those subregions that show a specific chromatin state change between the sample groups, a feature we refer to as “chromatin dynamics filtering”. To exemplify SCIDDO’s capabilities of identifying user-specified chromatin state changes, we designed a use case examining enhancers switching from an active to an inactive state (Section 4.3.8). To that end, EP300 peak datasets for HepG2 were downloaded from ENCODE (ENCFF674QCU and ENCFF806JJS) and merged using bedtools v2.26.0 [196]. To illustSCIDDO’s feature of “chromatin dynamics filtering”, chromatin states 7–11 (genic, active and weak enhancers) were considered as enhancer “on” states, and chromatin states 13, and 15–17 (heterochromatin, bivalent enhancer, and states of polycomb repression) were considered as enhancer “off” states.

4.2.7 Statistical background for SCIDDO

The theory behind the statistical evaluation available in SCIDDO has been developed in the context of biological sequence analysis, e.g., to identify runs of hydrophobic amino acids in protein sequences [117, 118]. Since the theory was left unaltered, we give only a compact overview to introduce the necessary concepts and nomenclature. The chromatin state map of each sample in the SCIDDO dataset can be represented as a sequence $X = \{x_1 \dots x_p \dots x_n\}$. Here, the x_p are assumed to be random variables over an alphabet A and n is the length of the sequence (more details are given in the listing below). In our case, $|A| = 18$ representing the 18 different chromatin states of the CMM18 model. Each pair of states (a^i, a^j) is assigned a score s^{ij} where $s^{ij} < 0$ indicates state similarity (regions with putatively consistent activity) and $s^{ij} > 0$ indicates state dissimilarity (regions with putatively differential activity; see below for derivation of the s^{ij}). We omit the superscript ij in the following to improve readability. When comparing two chromatin state maps X, Y, each state pairing (x_p, y_p) is assigned the respective score s as defined above. This results in a sequence of scores $S = \{s_1 \dots s_n\}$ that is scanned for subsegments of highest cumulative score. This approach is called local score computation and can be done efficiently with a linear time algorithm [208]. The set of all maximal scoring disjoint segments returned by this algorithm represents the set of candidate regions for the respective chromatin state map comparison. The (unnormalized) raw score R of a candidate region is simply defined as the sum over all scores in the candidate region $R = \sum_{k \leq p \leq l} s_p$ where k and l indicate the position of the leftmost and of the rightmost genomic bin included in the candidate region. These cumulative scores have to be normalized to account for the fact that higher scores have a higher chance of occurring with increasing sequence length. This normalization step requires the estimation of two statistical parameters λ and K (for detailed derivation of these parameters, see [118]). Since both λ and K lack a biologically meaningful interpretation, they can be simply thought of as scaling parameters for the scoring system and the search space. For this parameter estimation, SCIDDO relies on the routines implemented in BLAST v2.7.1 [8]. Additionally, four assumptions are needed for the theory to be applicable, which then allows for modeling the limiting behavior of the score distribution as Gumbel-type extreme value distribution (see [118]):

1. The sequences are infinitely long
2. The x_p are independent and identically distributed (i.i.d.) random variables
3. A positive score must be possible
4. The expected score is negative

Assumptions 1. and 2. of course do not apply to any biological sequence, but are needed for reasons of mathematical tractability beyond the scope of this thesis [118]. Assumptions 3. and 4. are tested by SCIDDO before starting the actual analysis, safeguarding against errors in the statistical evaluation. Under these conditions, the Expect value (E) for a DCD with raw score R is then calculated as

$$E = K \cdot L \cdot e^{-\lambda R} \tag{4.2}$$

where the factor L is the length of the chromosomal sequence adapted for replicate variation. Because SCIDDO has been designed to compare (small) groups of replicates against each other, we adapted the calculation of the total length of the sequence. For the sake of the argument, consider the ideal but unrealistic scenario of identical biological replicates in both sample groups. In this case, a SCIDDO analysis would always result in the same output, irrespective of the two samples being compared. Hence, intuitively, the number of (identical) replicates per group should not increase the statistical stringency, i.e., adding a perfect replicate to a group should not increase the sequence length factor L. Under more realistic conditions of imperfect but still high-quality biological replicates, SCIDDO only adds those positions to the total sequence length L that show a new chromatin state compared to all other replicates already in the group. A pseudocode for this computation is given as Algorithm 1 below:

Algorithm 1: Compute length normalization factor L

```
Input: Chromatin state maps for 2 sample groups (GROUPS), each consisting of at least
       one replicate $r$
Output: Length normalization factor $L_c$ per chromosome $c$

for $c \in CHROMOSOMES$ do
    $L_c = 0$
    $n = |c|$
    for $G \in GROUPS$ do
        $states_{1...n} = \emptyset$
        for $r \in G$ do
            $L_c = L_c + \sum_{p=1}^{n} \mathbb{1}(r_p \notin states_p)$
            for $p = 1$ to $n$ do
                $states_p = states_p \cup s_p$
            end
        end
    end
    $L_c = L_c - n$
end
```

4.2.8 Fit of random scores to Gumbel-type extreme value distribution

The calculation of the E-value as described above assumes a null model of random sequences. According to the theory (see Theorem 1 in [118] and examples in [116]) the normalized maximal scores should follow a Gumbel-type extreme value distribution when comparing random state sequences, in the limit of the sequence length n. Because SCIDDO supports the use of customized scoring schemes, it also supports the user in assessing if the chosen scoring scheme adheres to this theoretical assumption. To that end, SCIDDO scans the randomly shuffled chromatin state maps of all sample pairs for high scoring subsegments and retains only the maximally scoring subsegment per chromosome comparison; if several segments with identical scores emerge, only the first one is kept. This process is iterated until a pre-specified number of these “random” scores have been found. The user can then use these “random” scores and, e.g., assess their fit to a Gumbel-type extreme value distribution following our example (see Section 4.3.2).

4.2.9 Derivation of pairwise chromatin state similarity scores

The theoretical considerations presented in the previous sections do not require the use of sophisticated scoring schemes that are well-grounded in theory, e.g., rather simple “match/mismatch” or empirically derived scoring schemes can be used if considered appropriate [117]. We thus decided to use the emission probability vectors of the 18 chromatin states (= the hidden states of the ChromHMM Hidden Markov Model) to compute pairwise similarity scores. The state emissions $E_i = \left(e_i^h \ldots e_i^h\right)$ for state a_i represent a probability distribution over the observed outputs, i.e., over the observed six histone modifications h. Divergence measures are commonly employed to quantify the difference between probability distributions (see also Section 4.2.3), and in our case, it seems plausible that a symmetric measure should be used to compute the pairwise difference between chromatin states. This motivated using the symmetric Jensen-Shannon divergence (JSD) [143] to compute chromatin state similarities

$$JSD(E_i, E_j) = 2 \cdot H\left(\frac{E_i + E_j}{2}\right) - H\left(E_i\right) - H\left(E_j\right) \tag{4.3}$$

where H is the Shannon entropy

$$H(E_i) = -\sum_{h=1}^{6} e_i^h \cdot \log(e_i^h) \tag{4.4}$$

Because the JSD has a lower bound of 0, the pairwise similarities for each state were shifted by subtracting the mean JSD[5]. This resulted in negative scores for similar states (JSD near zero) and positive scores for dissimilar states. Scores are commonly represented by integer values, which we realized by multiplying the real-valued scores by a factor of 10 and rounding them to integers afterwards. Before starting the differential analysis, SCIDDO checks the adherence to assumptions 3. and 4. (Section 4.2.7) for any custom scoring scheme such as our JSD-derived one to ensure applicability of the statistics introduced in Section 4.2.7.

A peculiarity of chromatin state maps is the so-called background state (state 18 labeled as “quiescent” in the CMM18 model). This state represents the lack of any detectable signal in the input data. As it is *a priori* impossible to identify the true source for this lack of a signal, i.e., it could be a technical artifact or biologically meaningful, the background state needs to be handled with special care in the interpretation of chromatin state maps. We decided to implement a conservative strategy and replaced all pairwise state similarities involving the background state with the minimal score generated with our JSD-based approach. In other words, the background state is similar, i.e., not differential relative to all other chromatin states. We opted for this strategy to avoid finding differential chromatin domains that are dominated by the background state and could thus be challenging to interpret.

4.2.10 Code availability and study replication

The full source code of the SCIDDO command line tool is available under a GPLv3 license at github.com/ptrebert/sciddo.

The code for replicating all results presented in this chapter is publicly accessible under doi.org/10.17617/1.6K. All data preprocessing and analysis pipelines have been implemented in

[5] The decision to use the mean JSD for discriminating between similar and dissimilar chromatin states was driven by practical considerations. In the future, a more comprehensive understanding of chromatin differences between cell types may allow for a systematic approach based on, e.g., known regions of differential chromatin.

Python/Ruffus [90]. All figures except for Figure 4.1 can be recreated using the respective Jupyter Notebooks[6] available in the aforementioned repository.

4.3 Results

4.3.1 Score-based identification of differential chromatin domains

The differential analysis with SCIDDO consists of two major parts, data preparation and the actual analysis run (see Figure 4.1 for an overview). In the data preparation step (Figure 4.1 step (A)), SCIDDO creates a single coherent dataset storing all data and metadata relevant for the analysis to support later reproducibility of the results. As part of the data preparation, the state emission probabilities of the chromatin state segmentation model are used to compute pairwise chromatin state dissimilarities (see Section 4.2.9). Starting from this dataset, SCIDDO then performs the differential analysis as follows: for each comparison contrasting sample group X versus group Y, SCIDDO first compares individual replicates against each other, say, X-2 versus Y-1 (Figure 4.1 step (B)). In this process, each observed chromatin state pair in the two chromatin state maps is assigned a score that quantifies the dissimilarity of the two states: positive scores indicate state dissimilarity, and negative scores indicate state similarity (Figure 4.1 step (C) and Section 4.2.7). Candidate regions showing differential chromatin marking are identified on this level of replicate comparisons by searching for chromosomal segments that show a high cumulative score. The magnitude of these cumulative scores can be taken as an indicator of the chromatin state dissimilarity in the chromosomal segment; hence, we refer to this cumulative score as the differential chromatin score (DCS) of the segment (Figure 4.1 step (C) to (D)). It should be pointed out that extracting segments based on (locally) maximal DCSs implies also a maximization of the segment length, and no (predefined) minimum or maximum length has to be specified. To proceed from candidate regions identified in individual replicate comparisons (e.g., X-2 versus Y-1) to candidate regions that are assumed to be representative of all samples X versus Y, overlapping candidate regions are merged by averaging their DCSs and taking the union of their genomic coverages (Figure 4.1 step (E)). As the final step in the analysis, the segment DCSs are turned into an Expect (E) value, which allows for filtering the resulting candidate regions for their statistical significance (Figure 4.1 step (F)). The E-value (see Section 4.2.7) has the interpretation of indicating how many candidate regions with at least a similarly high DCS could arise simply due to chance when comparing random sequences of the same length. In other words, when filtering the candidate regions for a default E-value of less than 1 to call DCDs, SCIDDO restricts the results to those chromosomal regions where the chromatin states are so different between the samples that one would not expect to find such a difference simply due to chance. To simplify visualizations, we report E-values after a negative log10 transform in the remainder of this study. The aforementioned threshold of 1 is thus transformed to 0 and larger E-values indicate higher statistical stringency.

[6] http://jupyter.org

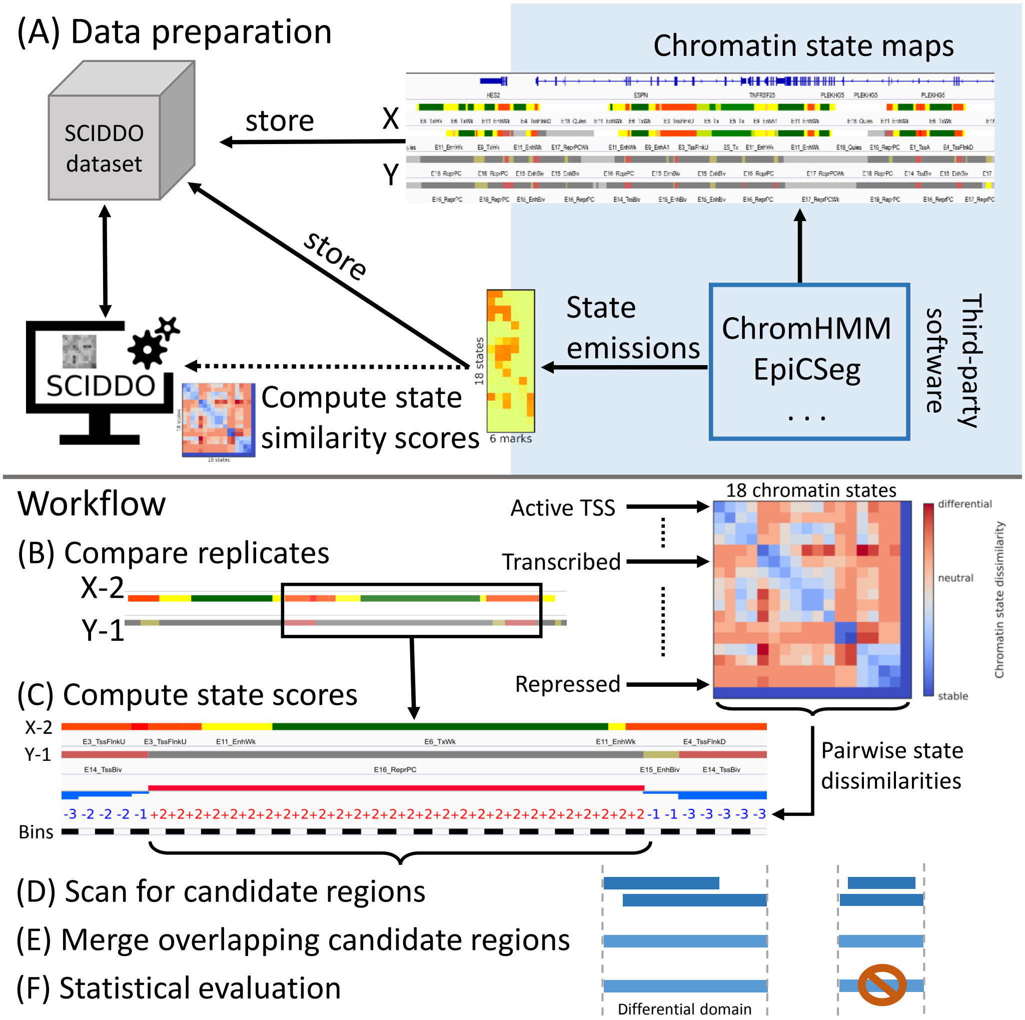


Figure 4.1: Overview of SCIDDO's workflow for identifying differential chromatin domains: (A) Data preparation: chromatin state maps can be generated using common tools (blue shaded area). The chromatin state maps for all replicates of sample groups X and Y are stored together with the chromatin state emission probabilities in a SCIDDO dataset to ensure later reproducibility of the analysis. The state emission probabilities are used to compute chromatin state similarity scores. For the analysis presented in this chapter, the CMM18 model was exclusively used to produce the input data. (B)–(F) Workflow: (B) the differential analysis starts by comparing all replicate pairs in the dataset, here exemplified as X-2 vs. Y-1. (C) All observed chromatin state pairs are scored with their respective dissimilarity score. (D) The resulting score sequences are scanned for high-scoring candidate regions. (E) Overlapping candidate regions of all replicate pairs are then merged and (F) filtered after statistical evaluation to generate the final set of differential chromatin domains.

We performed a differential analysis for all six possible pairings of sample groups in our dataset, i.e., (i) HepG2 vs. hepatocytes; (ii) HepG2 vs. monocytes; (iii) HepG2 vs. macrophages; (iv) hepatocytes vs. monocytes; (v) hepatocytes vs. macrophages, and (vi) monocytes vs. macrophages. Before executing the SCIDDO workflow, we confirmed that the generated chromatin state segmentation maps reflect the assumed relationships among all cell types by determining the fraction of identically assigned chromatin states between sample pairs (Figure B.2). This analysis result conforms to the expectation that monocytes and macrophages are considerably similar on the chromatin state level, whereas primary hepatocytes and HepG2 are not as similar (but still closer to each other than to the hematopoietic cell types).

Additionally, we tested our assumption that the predefined chromatin states of the CMM18 model capture relevant regulatory biology in our data despite the fact that these chromatin states were originally defined based on the REMC dataset. We trained and applied a new 18-state ChromHMM model (NEW18) on our dataset and found an acceptable agreement between the predefined CMM18 and the NEW18 chromatin states (Figure B.3, see Section 4.2.3 for details). Hence, we decided to take advantage of the complete annotation available for the REMC CMM18 model and limited our analyses to the CMM18 chromatin state maps.

The subsequent SCIDDO analysis including data preparation completed within minutes on a moderately powerful compute server (Table B.2). The results presented for this analysis are structured as follows: first, we provide some evidence that our data follow the theoretical assumptions necessary for the statistical evaluation. Next, we highlight the robustness of SCIDDO's results across replicates and then provide a more biology-oriented characterization of the identified DCDs.

4.3.2 Differential chromatin scores follow extreme value distribution

The last step in the SCIDDO workflow described above consists of turning the DCSs into an E-value that is used for filtering the set of candidate regions to obtain the final set of DCDs. This step relies on theory developed for biological sequence analysis (see Materials and Methods) and requires first a normalization of the raw cumulative DCSs to account for the fact that comparing longer chromosomal sequences increases the chances of observing higher cumulative DCSs. This normalization uses two estimated statistical parameters, λ and K that lack a biological interpretation, but can be thought of as scaling factors for the scoring system and the sequence length, respectively. Second, the theory assumes a null model of random sequences, and under this null model, the distribution of the scores should in the limit converge in distribution to a Gumbel-type extreme value distribution (Section 4.2.8). We confirmed that this is indeed the case in our analysis by comparing randomized chromatin state maps with each other and fitting all maximal DCSs identified during this sampling procedure to a Gumbel distribution (Figure 4.2 left panel). We also plotted the per-chromosome estimates of the statistical parameters λ and K that are needed for the score normalization (Figure 4.2 right panel), and could confirm that the estimates are within reasonable bounds given examples from literature [116]. The observed agreement with theory thus supports the last step in the SCIDDO analysis (Figure 4.1 step (F)) of filtering candidate regions based on their E-value.

4.3.3 SCIDDO robustly identifies differential chromatin domains

Histone ChIP-seq data is known to be affected by various sources of noise ranging from artifacts introduced during library preparation, to irregularities caused by varying mappability in the reference genome, or to spurious signal due to unspecific antibody binding (see Chapter 2, Section 2.3.1 for more details). In combination, biological and technical variation can render any differential analysis pointless if the results are dominated by noise, and not by the biological signal

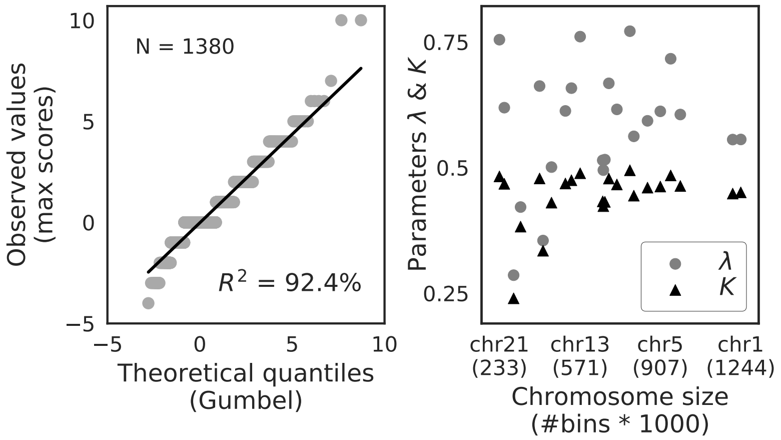


Figure 4.2: Observed maximal scores and parameter estimates follow theoretical assumptions (A) Probability plot of all normalized maximal scores derived from comparing random sequences (y-axis; see Section 4.2.8 for details) fit to the theoretical quantiles of a Gumbel-type extreme value distribution (x-axis). Here, we jointly fitted all "random" scores of all chromosomes to simplify the visualization. (B) Chromosomes are sorted by increasing size (in genomic bins) from left to right (x-axis) and the per-chromosome estimates of the two statistical parameters λ (gray points) and K (black triangles) are plotted on the same scale (y–axis). R^2: coefficient of determination

of interest. To test if the identified candidate regions were indeed representative and not replicate-specific, we computed the Spearman correlation of the E-values between all overlapping candidate regions. We visualized an exemplary case selected based on the mean of all comparisons. This exemplary case shows a Spearman correlation of 0.72 between the candidate regions (Figure 4.3). The red bars in the lower left corner indicate candidate regions that are unique to the respective replicate comparison. It can be observed that unique candidate regions tend to have comparatively lower E-values whereas those candidate regions found in both replicate comparisons tend to have higher E-values. In general, the average Spearman correlations across all replicate comparisons are consistently in high range from 0.67 (HepG2 vs. hepatocytes) to 0.73 (HepG2 vs. monocytes; see Table B.3).

4.3.4 Differential chromatin domains occur in various regulatory contexts

Since it is well-established that histone marks occur in various regulatory contexts, e.g., ranging from promoters and enhancers to gene bodies, it stands to reason that *bona fide* DCDs should predominantly occur in similar regulatory contexts. To test this hypothesis, we intersected the DCDs identified by SCIDDO with various annotation datasets and observed that, in general, around 80 to 90% of all DCDs overlap with at least one type of genomic annotation (Figure 4.4). Since there is no theory that would enable us to formulate an *a priori* expectation about the extent to which differences on the chromatin level should occur between any two cell types, we cannot assess the plausibility of the absolute numbers of identified domains. Nevertheless, it can be observed that the lowest number of domains is detected in the comparison of monocytes to macrophages (Figure 4.4F), i.e., when comparing the two most closely related cell types in our dataset. For all other

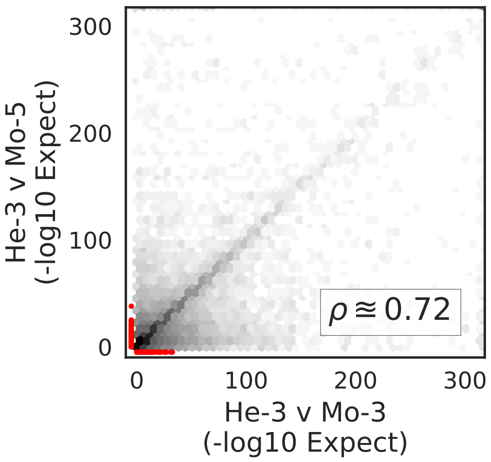


Figure 4.3: Candidate regions are robustly identified across individual replicates: exemplified agreement of candidate regions identified in replicate comparisons. E-values of candidate regions identified for He-3 vs. Mo-3 (x-axis) are plotted against E–values of overlapping candidate regions identified for He-3 vs. Mo-5 (y-axis). The red area indicates E-values of those candidate regions that are unique to the respective replicate comparison. ρ: Spearman correlation of E-values

comparisons, the number of identified chromatin domains is approximately four- to more than five-fold higher, but yet shows a similar tendency of a smaller number of identified chromatin domains for more closely related cell types.

These results also illustrate that the distribution of overlaps seems not to be affected by the number of DCDs identified. In all comparisons, at least approximately 70% of the DCDs overlap with at least one regulatory region annotated in the Ensembl Regulatory Build [266]. The Regulatory Build comprises different types of regulatory regions and has extensive genome coverage. Hence, the Regulatory Build enables us to interpret the relevance of DCDs in light of various functional categories. Since the distribution of genomic locations of the DCDs seems fairly similar across all comparisons, and analogous observations can be made when examining the length distribution of the DCDs (Figure B.4), we examined if there is a difference in DCD E-values aggregated over all comparisons (Figure 4.5). DCDs overlapping any regulatory region show higher E-values compared to those DCDs that have no overlaps (Figure 4.5, bottom panel). This effect is most pronounced for annotated promoters and transcription factor binding sites (TFBS), and this seems not to be an effect of regulatory region size (Figure 4.5, top panel). The average number of distinct regulatory region overlaps per DCD shows that a DCD often spans several of the shorter regulatory regions, with the exception of TFBS, which is the least abundant region type with a median size $<$ 1 kbp in the Regulatory Build. At the other end of the size spectrum are promoters, which also show hardly any variation around a median of one DCD overlap per promoter.

4.3.5 Formation of differential chromatin domains affects gene expression

The results presented in the previous section indicate that DCDs largely overlap with a variety of regulatory regions, and thus it seems plausible that the formation of a DCD should have functional consequences, e.g., by modulating gene expression levels. Apart from basic considerations

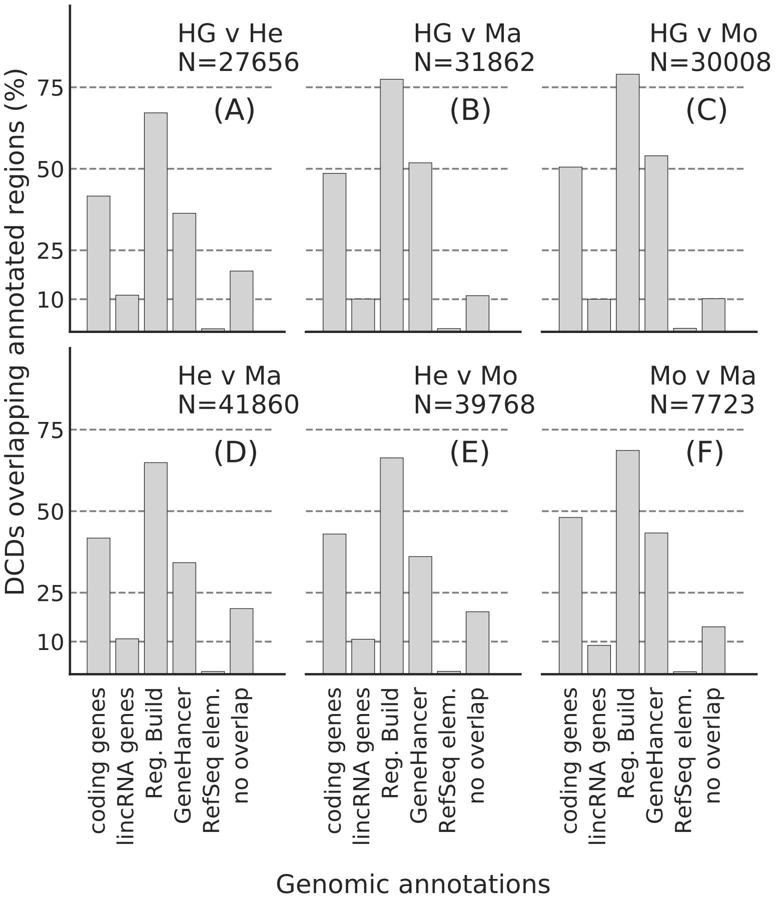


Figure 4.4: Differential chromatin domains overlap with annotated regulatory regions: (A)–(F) bar heights indicate percentage of identified differential chromatin domains that overlap with different genomic annotations for all six sample group comparisons. HG: HepG2; He: hepatocytes; Ma: macrophages; Mo: monocytes; N: total number of identified DCDs; coding genes: Gencode v21 protein-coding genes; lincRNA genes: Gencode v21 lincRNA genes; Reg. Build: Ensembl Regulatory Build v78; GeneHancer: GeneHancer annotated enhancers limited to Gencode v21 gene set; Refseq elem.: RefSeq functional elements

about the magnitude of the observed E-values, we also hypothesized that DCDs covering larger parts of the gene body could indicate stronger changes in gene expression. To give a canonical example, a gene that is entirely repressed by means of polycomb-mediated silencing should be enriched for the histone mark H3K27me3, and this marking should be replaced by H3K36me3 as soon as the gene is activated and actively transcribed (see Chapter 2, Section 2.1.1.4). On the other hand, if the gene expression is modulated, e.g., by changing transcription factor binding in enhancer regions, the effect on the chromatin marking in the gene body could arguably be less pronounced. To investigate this hypothesis, we stratified all genes by the fraction of their gene body length being covered by a DCD (no overlap in gene body or enhancers, less or more than 50% gene body overlap). Next, we computed gene expression fold changes using DESeq2 [147] (Section 4.2.4) for the six sample group comparisons and visualized the fold change for all genes in the three DCD overlap groups as a cumulative distribution (Figures 4.6 and B.5). The curves indicate that genes covered by more than 50% of their body length with a DCD indeed exhibit

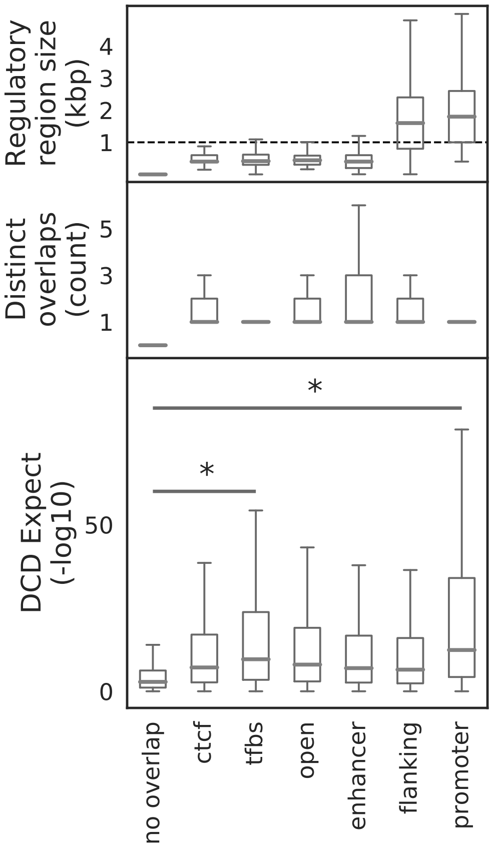


Figure 4.5: E-value distribution of DCDs overlapping regulatory regions: (bottom) box plots show distribution of E-values of all differential chromatin domains overlapping regulatory region types as annotated in the Ensembl Regulatory Build (v78) aggregated over all sample comparisons. As an example, differences in magnitude of E-values between the groups "tfbs vs. no overlap" and "promoter vs. no overlap" were assessed with a two-sided Mann-Whitney-U test and considered significant "*" at $p < 0.01$. (middle) box plots show distinct overlaps per DCD, i.e., the number of regulatory regions of that type overlapping the same DCD. (top) box plots show size distribution of the Ensembl regulatory regions. Dashed line indicates a size of 1,000 bp. Regulatory region types: ctcf: CTCF binding sites; tfbs: transcription factor binding sites; open: regions of open chromatin; enhancer: enhancer; flanking: promoter-flanking regions; promoter: promoter

stronger changes in their expression level (orange lines). A similar albeit weaker effect can be observed for genes having less than 50% of their body or their promoter covered by a DCD (blue lines). In many cases, the difference in fold change relative to the group of genes that does not overlap a DCD is significant. Additionally, we applied the same method to test if the number of gene-associated enhancers that overlap a DCD had a similar bearing on gene expression (Figure 4.6 and B.5, middle and right panels). This enhancer-centric view shows a stable pattern across most sample comparisons that indicates that stronger changes in gene expression occur if more gene-associated enhancers overlap a DCD. This observation is particularly intriguing when restricting the view on intergenic enhancers, where, as opposed to intragenic enhancers, there is lower chance of a coincidental overlap with a DCD. In general, a small but noticeable difference compared to the no-DCD overlap group (gray dashed line) can be expected as soon as 2–3 enhancers show a DCD (magenta curve).

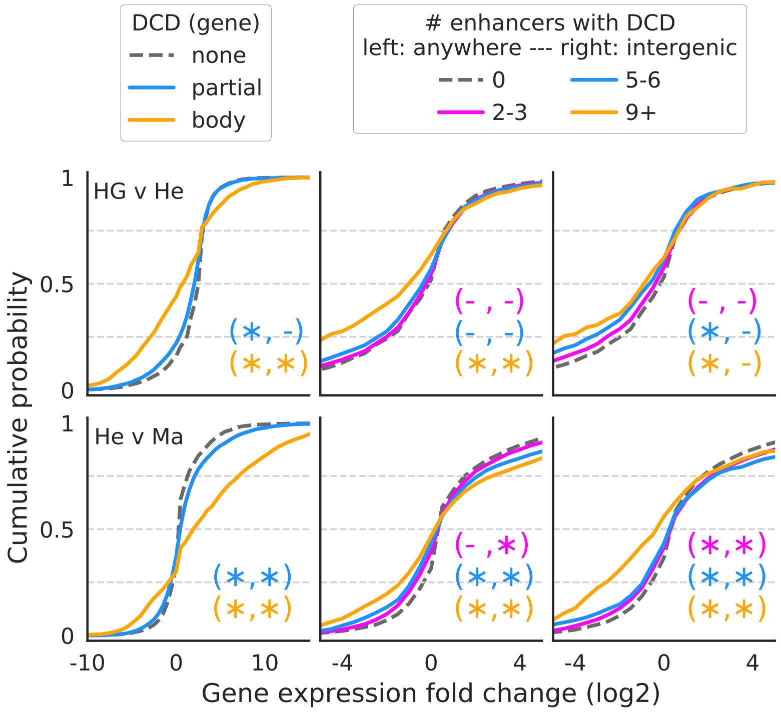


Figure 4.6: DCDs overlapping gene bodies and enhancers affect gene expression: (left) genes were stratified by the amount of DCD overlap either covering more than 50% of the body (body; orange curve) or less than 50% of the body or the promoter region (partial; blue curve). Expression fold change of the genes in the respective groups is plotted along the x-axis within a restricted window for improved readability. Statistical significance of the difference in mean fold change of the groups relative to the no overlap group ("none", gray dashed line) was computed separately for negative and positive fold change genes using a two-sided Mann-Whitney-U test ("*" significant at $p < 0.01$, "-" not significant otherwise). (middle and right) the same analysis as for the gene body was performed, but here counting the number of intra- and intergenic enhancers (anywhere, middle) or only intergenic enhancers (right) per gene that overlap a DCD. Expression fold changes plotted within a restricted window for improved readability. Statistical significance assessed as before.

4.3.6 SCIDDO detects chromatin changes in differentially expressed genes

By design, SCIDDO does not impose any restrictions on the regions of interest that can be interrogated in a differential analysis. Since there is no general model of chromatin variation that would enable us to assess the plausibility of the identified differential chromatin domains irrespective of their genomic context, we decided to focus on a small-scale case study that is arguably of broad biological interest.

We investigated to what extent DCDs can be used to specifically identify DEGs. As ground truth for this analysis, we used the same DESeq2 results as above, but applied a threshold to split the genes into differentially expressed and stable ones (Section 4.2.4). As a first step, we checked what percentage of DEGs could be recovered using the DCDs identified by SCIDDO (Figure 4.7). For four out of the six sample comparisons, more than 90% of all DEGs could be recovered with DCDs either overlapping the gene body, the gene promoter or at least one gene-associated

enhancer. For the comparison of HepG2 to primary hepatocytes (Figure 4.7A), approximately 81% of DEGs could be recovered, and for the comparison of monocytes to macrophages, 54% of all DEGs were recoverable by using DCDs (Figure 4.7F). The comparatively lower rate of DEG recovery for the monocyte to macrophage analysis seems to be in line with the already observed trend of fewer differences on the chromatin level with increasing cellular relatedness (e.g, see Figure 4.4). We present a more in-depth analysis of this observation in Section 4.3.7.
Next, we tested if it was possible to broadly distinguish between DEGs and stably expressed genes by thresholding on the E-values of the DCDs that overlap gene bodies. To that end, we stratified the set of DEGs based on their fold change into three groups (top 20%, middle and bottom 40%) and plotted the E-value distribution of the DCDs for these three groups and for all other chromatin domains (Figure 4.8, bottom panel). We find that the top 20% of all DEGs overlap DCDs that have a significantly higher E-value on average relative to DCDs overlapping the remaining DEGs. Furthermore, it is interesting to observe that the E-value distribution of the DCDs overlapping stable genes is similar to those that do not overlap any gene (but could, e.g., overlap with intergenic enhancers). The number of distinct DCDs that overlap any given gene shows no notable variation across all groups (Figure 4.8, middle panel). The distribution of the gene body lengths in the respective groups appears to be fairly balanced (Figure 4.8, top panel) and thus does not suggest that the number of DCD overlaps or the observed difference in E-value distribution is a simple effect of gene body length. We explicitly confirmed this by repeating the analysis, but this time stratifying DEGs by gene body length (Figure B.6). The E-values of the DCDs overlapping the longest genes are comparatively lower, and this suggests that larger E-values are probably not a result of increasing gene body length.

4.3.7 Methodological and biological limitations for chromatin–based detection of differentially expressed genes

The theory borrowed from local scoring and implemented in SCIDDO is used to assign a measure of statistical stringency — the E-value — to each discovered DCD. Yet, the theory does not offer a way to decide what threshold on the E-value best separates genuine from chance observations. The necessary normalization to account for the length of the sequences being compared immediately suggests that short but biologically meaningful differential regions will be assigned an (untransformed) E-value above SCIDDO's default threshold of 1.
We checked the extent to which the default E-value threshold of 1 could limit SCIDDO's ability to identify — especially short — DEGs. We binned all DEGs by their gene body length and plotted the amount of genes with a DCD overlapping their gene body at E-value thresholds of 1 and 100 (Figure 4.9). The histogram shows the expected behavior of SCIDDO to predominantly recover longer DEGs by means of identifying a DCD in their gene body. However, relaxing the E-value threshold seems not to affect this general trend as the additional DEGs also show a tendency toward longer gene bodies. We thus wondered if other technical or biological artifacts might exacerbate the detection of DEGs on the chromatin level. We focused specifically on the comparison of monocytes to macrophages where approximately only 54% of all DEGs could be

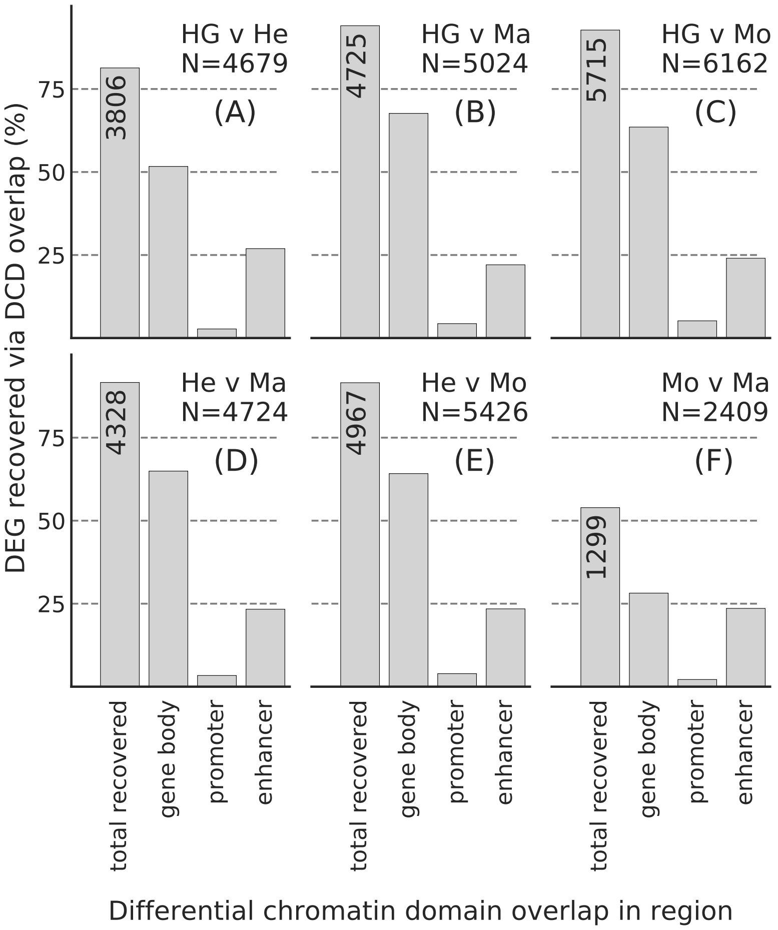


Figure 4.7: Differential chromatin domains recover differentially expressed genes: (A)–(F) bar heights indicate percentage of recovered differentially expressed genes by counting overlaps with differential chromatin domains in gene bodies, in gene promoters (but not in gene bodies) or in gene-associated enhancers (but not in gene bodies or gene promoters). The leftmost bar is annotated with the total number of recovered genes. N: total number of differentially expressed genes per comparison.

recovered using DCDs (see Figure 4.7F).

We examined if artifacts in the data could be the reason for the low DEG recovery rate. Besides chromatin states with annotated function, chromatin state maps usually include a so-called background state that represents regions of no detectable signal (state number 18 labeled as “quiescent” in the CMM18 model). It is important to realize, though, that the interpretation of this background state is difficult. It is conceivable that technical problems caused this lack of a signal in certain regions of the genome, but the lacking signal may be biologically meaningful in other regions. Moreover, the six canonical histone marks included in this study certainly cover a wide range of functionally important chromatin signals (Chapter 2, Section 2.1.1.4), but they do not represent the entire regulatory chromatin landscape. For example, the recently characterized H3K122ac histone modification is also found at active enhancers that lack the canonical H3K27ac marking [191]. Given these uncertainties, we opted for a conservative approach and considered

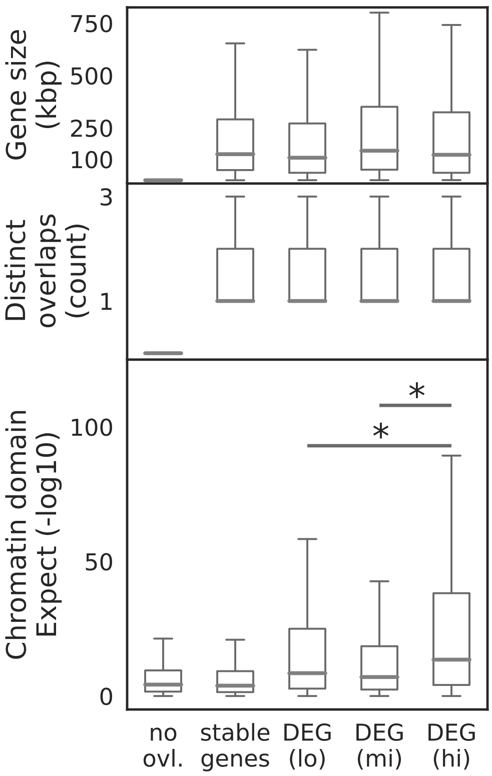


Figure 4.8: E-value distribution of DCDs overlapping gene bodies: genes were stratified into four groups based on their expression fold change (stable/no change, lowest 40% [lo], middle 40% [mi], and top 20% [hi] of DEGs according to expression fold change). (bottom) box plots show distribution of E-values of all DCDs overlapping gene bodies in the respective groups aggregated over all sample comparisons. The no overlap group (no ovl.) contains all E-values of DCDs not overlapping any gene. Differences in magnitude of E-values for the two comparisons “lo vs. hi” and “mi vs. hi” were assessed with a two-sided Mann-Whitney-U test and considered significant “*” at $p < 0.01$. (middle) box plots show number of distinct DCD overlaps per gene. (top) box plots show gene body length distribution of all genes in the respective group.

the background state as not differential relative to all other chromatin states (see Section 4.2.9). We evaluated how many DEGs might not be recoverable under these conditions for the monocyte to macrophage comparison. For each of the 1,110 DEGs that could not be recovered, we computed the percentage of the gene body length covered with the background state (averaged over all replicates in the respective groups). We found that close to a hundred genes that are covered to at least 60% with the background state are shared between the monocyte and the macrophage group (Figure 4.10A). At a higher threshold of 80% body coverage, this number drops to 35 genes. Given that this considers genes that are in the same uninformative chromatin state to roughly the same extent in all samples — and being differentially expressed at the same time — it seems justifiable to assume that the non-detection of these genes is not a limitation of SCIDDO. When focusing on the genes that are covered with the background state in either monocytes or macrophages, the numbers rise considerably (Figure 4.10B). For the lower threshold of at least 60% coverage, 164 genes show the uninformative chromatin state in their gene body, and when raising the threshold to at least 80% coverage, 72 genes are still affected. In this scenario, the

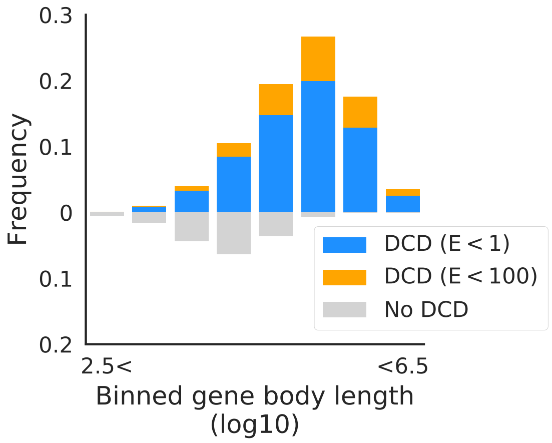


Figure 4.9: Relaxing E-value threshold does not improve detection of short DEGs: all DEGs for all six comparisons were binned based on their gene body size (x-axis) and classified based on overlapping DCDs in their gene body (y-axis). DCDs were called with the default threshold of E $<$ 1 (blue) and with a relaxed threshold of E $<$ 100 (orange).

non-detection of the DEGs is hence largely driven by the lack of a signal in one of the two sample groups.

Considerations involving the background state might explain a few hundred cases of DEGs that could not be recovered by SCIDDO. This implies, however, that a considerable amount of DEGs were assigned biologically meaningful chromatin states and yet were not detectable by SCIDDO.

We hypothesized that a plausible cause for this could be a comparatively weak change in gene expression for non-detectable genes. When a gene is switched from “off” to “on”, a substantial change in the histone marking can be expected. However, if the gene is already actively transcribed and then simply upregulated, e.g., by activating additional enhancer elements (see Figures 4.6 and B.5), it is not obvious why this change in expression should lead to differential chromatin marking in the gene body. We tested this hypothesis by plotting the mean difference in expression, plus the minimal and maximal expression level in any sample, for all DEGs in the monocyte to macrophage comparison (Figure 4.10 (C)–(E)). We split the genes into three groups based on DCD overlap in their gene body, in any associated enhancers but not in the body, and no DCD overlap at all, i.e., the non-detectable genes. The mean change in gene expression is significantly higher in genes overlapping with a DCD compared to those genes that have no differential chromatin marking. Interestingly, the minimal expression level (Figure 4.10D) is still relatively high for those genes that show differential chromatin marking only in their enhancers. When relating the minimal to the maximal expression level (Figure 4.10D/E), the change in expression can be characterized as follows: genes with a DCD in their gene body jump from a low to a high expression level; genes with no DCD in their body but in their enhancer(s) show increased expression relative to an already high level, and genes with no DCD at all remain at a low to mildly elevated expression level. It should be pointed out that the implied directionality is supported by the observed expression changes for the monocyte to macrophage comparison

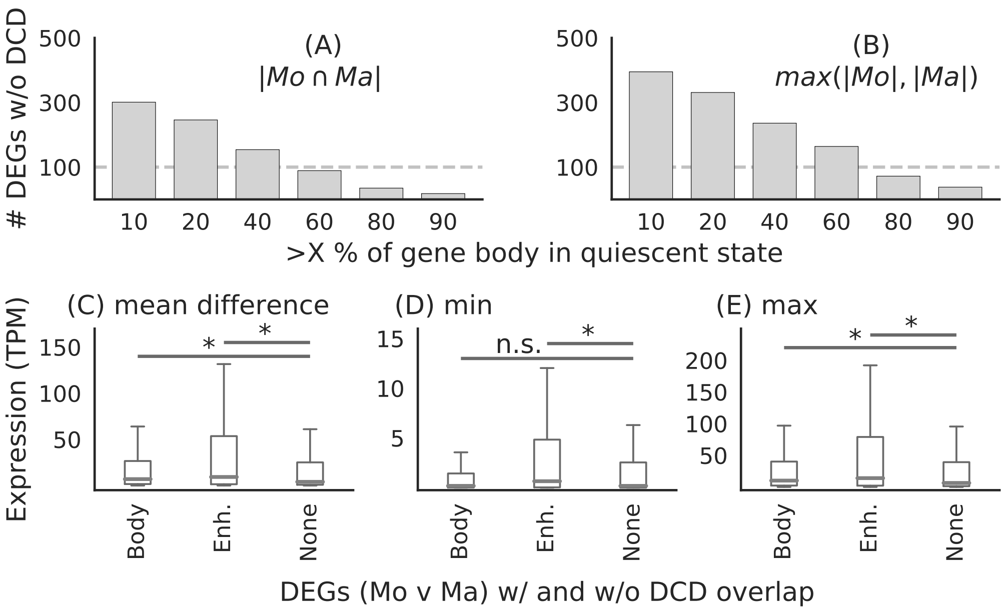


Figure 4.10: Uninformative chromatin state in gene bodies and moderate changes in expression complicate DEG recovery: (top) bar charts show DEGs binned according to the fraction of their gene body covered with the background "quiescent" chromatin state (x-axis). (A) height of bars depicts number of genes in intersection between monocyte (Mo) and macrophage (Ma) samples. (B) height of bars depicts maximal number of genes either from monocyte or from macrophage samples. (bottom) DEGs were stratified according to DCD overlap in gene body/promoter (Body), or in at least one enhancer (Enh.) or no DCD overlap (None). (C) box plots show distribution of gene expression values for absolute mean differences between monocyte and macrophage samples, and (D) for minimal expression, and (E) for maximal expression in any sample. Differences in magnitude were assessed using a two-sided Mann-Whitney-U test and considered significant "*" at $p < 0.01$ and not significant (n.s.) otherwise.

(Figure B.5a).

There is a multitude of mechanisms beyond the chromatin level that can fine-tune gene expression [51, 125, 178]. Given that the DEGs lacking any sign of differential chromatin marking show also limited dynamics in their expression changes, we wondered whether there was any evidence of post-transcriptional control of these genes. As control group, we selected all genes that were not classified as differentially expressed but nevertheless showed signs of differential chromatin marking in their gene body (N=760 for the monocyte to macrophage comparison). We then plotted the number of annotated micro RNA targets using the TargetScan v7.2 [3] annotation for both groups of genes (Figure B.7, bottom panel). There is indeed a small but statistically significant difference in the number of annotated micro RNA targets per gene between the two groups. This difference seems not to be caused by a difference in 3'-UTR length, where it is the group of DEGs without an overlapping DCD that has the larger 3'-UTR regions on average (Figure B.7, top panel).

4.3.8 SCIDDO affords direct interrogation of chromatin dynamics

A noteworthy feature of SCIDDO is the possibility to restrict the set of all DCDs to certain subregions based on the observed chromatin dynamics (see Section 4.3.8). Given that chromatin states generated by the CMM18 model have been assigned meaningful labels (Figure B.1 and Table B.1), users can exploit this easily interpretable annotation to perform this postprocessing step. We used this feature in combination with external validation data to investigate if it is possible to identify enhancers that switch from an "on" to an "off" state between two cell types. To this end, we selected two sets of chromatin state labels as representing active and inactive enhancer states (see Section 4.3.8). SCIDDO then uses these state labels to identify those subregions of a DCD where the chromatin change of interest can be observed between the selected cell types. It should be emphasized that, while the chromatin dynamics filtering is based on the identified DCDs, the individual subregions returned by SCIDDO cannot be statistically evaluated by computing an E-value. Subregions of a DCD can be as short as one or two genomic bins and, thus, the computed E-value of a subregion is unlikely to indicate statistical significance. For comparison, we downloaded several ENCODE peak datasets of the transcriptional co-activator EP300 (p300) for the cell line HepG2 (Section 4.3.8). Although EP300 is known to be highly predictive of tissue-specific enhancer activity [250], it cannot be assumed that all downloaded EP300 peaks mark active enhancers that are unique to HepG2, and are hence inactive in any other cell type. As a consequence, an exhaustive overlap between EP300 peaks and (switching) enhancer regions in DCDs cannot be expected. Instead, we hypothesized that it is more realistic to assume that any biologically meaningful enhancer switch within a DCD subregion should likely also show a change in EP300 occupancy. We investigated this hypothesis by plotting the count of EP300 peaks and their signal strength for all peaks generally overlapping DCDs, and for all peaks overlapping with DCD subregions showing enhancer switches from "on" to "off" and vice versa from "off" to "on" for the comparison of HepG2 to monocytes (Figure 4.11). There is a prominent difference both in absolute number of peaks and in signal strength for the two directions of enhancer switching. This example illustrates that SCIDDO can also offer support in downstream analysis by quickly identifying regions of specific and directed changes on the chromatin level.

4.3.9 Differential chromatin domains recover differentially expressed genes with increased stability compared to individual histone marks

The number of available tools that use chromatin state maps as input for a differential analysis is limited. ChromDet [39] is designed for group comparisons with at least 5 to 10 replicates each (personal communication between Marcel H. Schulz and Alfonso Valencia, ECCB 2018; see also footnote 1 on page 59), and thus did not give results on our dataset. Similarly, ChromDiff [264] could not identify any differential chromatin marking (in genes), presumably due to lacking statistical power given the limited number of replicates in our dataset. The Chromswitch package [111] can only process one chromatin state at a time, which complicates direct and fair comparisons with the DCDs identified by SCIDDO.

We thus decided to compare SCIDDO to PePr [267], an established tool for the differential analysis

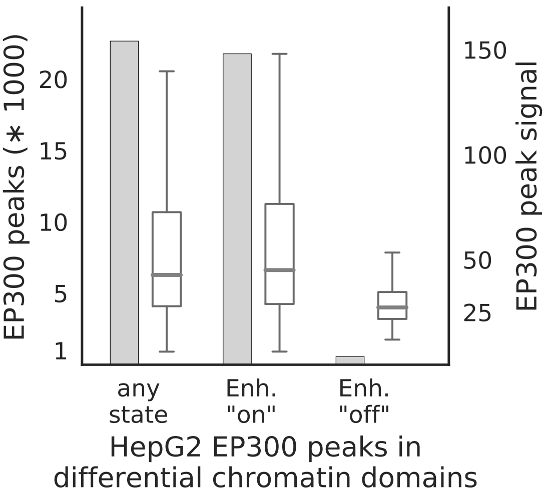


Figure 4.11: Chromatin dynamics at HepG2 enhancer elements: height of the bars depicts total number of peaks overlapping DCDs (left y-axis) and box plots show distribution of the signal of the overlapping EP300 peaks (right y-axis). The three groups represent (left) EP300 peaks overlapping with DCDs in general; (middle) with DCDs restricted to genomic locations showing an enhancer "on" state in HepG2; (right) with DCDs restricted to genomic locations showing an enhancer "off" state in HepG2. For all three groups, the DCDs identified in the HepG2 to monocyte comparison were used.

of individual histone marks that can process replicated samples. This strategy has the advantage of reflecting the canonical "rule-based" approach of interpreting histone marks in well-characterized regulatory contexts, e.g., by determining enhancer activity based on the presence of H3K27ac peaks (see Chapter 2, Section 2.1.1.4). Specifically, we used PePr to perform a differential analysis for the same six sample group comparisons and evaluated PePr's and SCIDDO's performance for the task of detecting DEGs based on differential chromatin marking (Section 4.2.5). To this end, we considered two different scenarios: first, genes overlapping at least one differential chromatin domain (SCIDDO) or having at least one H3K36me3 peak in one cell type but none in the other cell type (PePr) were labeled as differentially expressed. This strategy could be applied to all 20,091 genes in our gene annotation (gene set G1). In the second scenario, differential chromatin in gene bodies was taken into account in the same way, but as an additional requirement, at least three annotated enhancers of a gene had to show differential chromatin marking (H3K27ac peaks for PePr) to label the gene as differentially expressed. This reduced the number of genes in the evaluation set to 17,735 (88.3%; gene set G2), i.e., all genes that had at least three enhancers annotated. We compared the chromatin-based labeling of genes in sets G1 and G2 with the ground truth DEG labeling derived with DESeq2 [147] (Section 4.2.4). Although we settled for a fixed threshold on gene expression fold change (> 2) and p-value (< 0.01) to identify DEGs throughout this study, we varied these values for the comparison between SCIDDO and PePr to examine the stability of their performance for different levels of differential expression stringency. We calculated accuracy and F1 score for all sample comparisons and the gene expression fold changes 0.5, 1, 2 and 4 and p-values 0.1, 0.05, 0.01 and 0.001 for the two gene sets G1 and G2 (Figures 4.12 and B.8). In summary, SCIDDO's performance is superior to PePr. Averaged over all comparisons, SCIDDO shows an accuracy of 64.6% (G1) and 69.2% (G2) and a F1 score of 57.5% (G1) and 59.1% (G2) for the two different strategies of labeling a gene as differentially expressed. For PePr, the average performance scores are 57.6% (G1) and 57.7% (G2) accuracy and 54.6% (G1) and 54.7% (G2) F1 score.

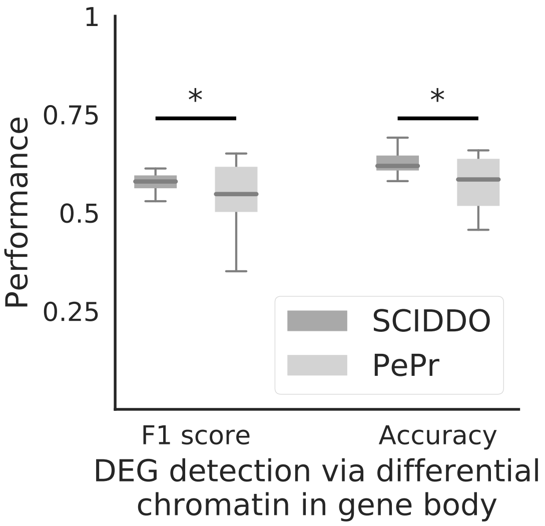


Figure 4.12: SCIDDO shows more stable performance at detecting DEGs (G1): box plots depict SCIDDO's and PePr's (light grey) performance of detecting DEGs quantified as F1 score (left) and as accuracy (right). Performance values are summarized over all sample group comparisons and for different thresholds on gene expression fold change (0.5, 1, 2 and 4) and on adjusted p-values (0.1, 0.05, 0.01 and 0.001) computed with DESeq2 to call DEGs. At least one DCD/differential H3K36me3 peak (PePr) was required in the gene body of a DEG to be considered detected on the chromatin level. Differences in performance were assessed with a one-sided Mann-Whitney-U test and considered significant "*" at $p < 0.01$.

4.4 Discussion

The use of chromatin state segmentation maps for large-scale annotation and interpretation of reference epigenomes is well-established in the field of computational epigenomics (see, e.g., [71, 72]). Nevertheless, comparatively little effort has been invested in the development of generally applicable software that assists researchers in exploiting these resources. To fill that gap, we developed SCIDDO, a new tool that implements a score-based approach for the fast detection of differential chromatin domains between potentially small groups of replicated samples.

The results presented above indicate that SCIDDO is able to robustly identify consistent sets of differential chromatin candidate regions across individual biological replicate comparisons. This observation suggests that SCIDDO is well-equipped for the commonly encountered situation of limited replicate availability while still offering a statistically sound evaluation of the detected DCDs. Although the statistics implemented in SCIDDO do not afford a theory-driven evaluation of the detected DCDs, e.g., no suitable E-value threshold is motivated by the theory, we could validate our findings in several biologically meaningful ways. The considerable overlap between the detected DCDs and various regulatory annotation datasets (Figure 4.4) suggests a functional role for the identified DCDs that is in line with published studies [12, 101, 165]. By relating gene-expression fold changes to DCD formation in gene bodies and gene-associated enhancers, we could show that this presumed functional role seems to have a measurable effect on gene expression behavior (Figures 4.6 and B.5). Our findings conform to the established view that extensive chromatin changes in gene bodies and in gene-associated enhancers are good indicators of the expected gene-expression fold change [115, 132, 188]. It should be emphasized that SCIDDO realizes this view on the interplay between chromatin changes and altered gene expression without directly quantifying differences on, e.g., the read count level. Nevertheless, SCIDDO is able to detect most DEGs (Figure 4.7), and shows a performance in such tasks that is on average superior and more stable compared to competing approaches, which implement more time-intensive strategies to differential chromatin analysis (Figure 4.12).

An observable trend in the dataset we analyzed is the limited variation on the chromatin level with increasing cellular relatedness, e.g., what we have detailed for the monocyte to macrophage comparison. Although this inverse relationship is plausible, it implies that there is a natural limit in “resolution” of differential chromatin state analyses that governs SCIDDO’s applicability in discerning cellular phenotypes or characterizing differentiation pathways. Although we did not investigate these potential limitations in depth, we collected multiple lines of evidence that illustrate various ways of how gene expression changes, and thus different cellular phenotypes, could be realized without necessarily leaving a detectable trace on the chromatin level (Figures 4.10 and B.7). One of these blind spots in chromatin state maps is the “quiescent” background state, i.e., the chromatin state without any detectable signal. If possible, a more fine-grained characterization of the background state would be a promising way of extending score-based differential chromatin analyses to cover even more regions of the (epi-) genome. For example, a widespread background state in gene bodies in only one sample group might be interpreted as biologically meaningful (Figure 4.10B), and thus, an adapted scoring for the background state in this context could plausibly increase DEG recovery rates via DCD overlap.
In total, the evidence supports the conclusion that SCIDDO’s score-based approach to differential chromatin analysis discovers biologically meaningful and interpretable DCDs. SCIDDO could thus become a useful software for chromatin bioinformatics and complement existing tools in analysis settings as described in this chapter.

CHAPTER 5

Epigenome-based Prediction of Gene Expression across Species

Lead-in This chapter is concerned with the following question: can bioinformatics bridge the data gap between model and non-model organisms in epigenomics? Section 5.1 provides a short motivation and an overview of relevant observations in previous cross-species epigenome studies. Based on these reports, we developed a bioinformatics pipeline tailored to perform an exploratory analysis across species boundaries. In Section 5.3, we present evidence indicating that there is indeed a case to be made for pure *in silico* approaches in data-scarce settings such as cross-species epigenomics involving non-model organisms. This chapter concludes with a discussion of the results presented in Section 5.4, and ideas for future work are included in Chapter 6.

The work presented in this chapter is an extended version of the manuscript Ebert et al. [65] (for details and author contributions, see Appendix E.1.3).

5.1 Background

Cross-species genome analysis is widely used for investigating evolutionary processes, identifying regulatory elements, improving genomic annotations, and studying the mechanisms underlying human diseases [28, 32, 129, 139, 161, 248, 258]. Recent progress with epigenome profiling technology has added a new dimension to genome comparisons. Cross-species comparisons that incorporate epigenomics and functional genomics data have opened up new ways of examining evolutionary processes, looking beyond genomic sequence conservation [249, 261, 263]. However, due to the high cost of generating epigenome data, as well as the cell-type specific and dynamic nature of epigenomic marks (Chapter 2), current reference epigenome datasets have been limited to a handful of species, most notably human and mouse [1, 48, 231, 241]. Genome sequencing efforts for other vertebrate species rarely include epigenome profiling [68, 128]. This has hampered the investigation of epigenome regulation in non-model organisms and precluded systematic cross-species epigenome analyses, with a few notable exceptions [146, 214, 249, 261].

There are several noteworthy observations in the aforementioned exceptions that bear some

relevance in the context of this chapter. Long et al. [146] investigated why computational predictions of CGIs in cold-blooded vertebrates (e.g., in zebrafish and frog) often do not overlap with gene promoters, whereas the inverse is the common case in (warm-blooded) mammals (e.g., in human and mouse). After experimentally determining non-methylated islands (NMIs)[1] in seven diverse vertebrate species, Long et al. discovered that the computational predictions of CGIs — taken as proxy for NMIs — were in strong disagreement with the experimentally determined NMIs. The experimentally determined NMIs showed the expected location pattern in or around gene promoters. After examining the nucleotide compositions of all seven species, Long et al. concluded that differences in CpG and G+C sequence content complicate the computational identification of NMIs in other (non-model) species. In other words, computational approaches to cross-species DNA methylome analysis should be carefully evaluated for their applicability if non-model species are included in the study.
Similar species-specific differences were reported by Schmidt et al. [214], who investigated TF binding of two liver-specific factors using ChIP-seq in five vertebrate species. The central findings by Schmidt et al. can be summarized as follows: most TF binding events are species-specific, and ultraconserved binding events, i.e., binding events shared among all five vertebrates, are rarely observed *in vivo*. Moreover, Schmidt et al. found that approximately only 50% of binding events lost in one species but present in at least two others could potentially be recovered by nearby TF turnover[2] events, suggesting substantial species-specific changes in the local regulatory environment. Schmidt et al. also made the important observation that there is no cross-species correlation between TFBS motif and binding event conservation, or between TF binding strength and binding event conservation. Given that Schmidt et al. selected liver because it is a functionally conserved organ in the five investigated species, their results seem to suggest that the regulatory landscape molded by TF binding is too highly dynamic to be amenable for accurate computational predictions across species if there is no explicit model available to account for the rapid TF turnover.

On the (histone) chromatin level, previous reports in the literature suggest a rather complex relation between the genome and the epigenome over evolutionary timescales. For example, Xiao et al. [261] examined several histone marks in cells from human, mouse and pig. Xiao et al. reported a notable tendency for epigenetic variation to be larger across species (interspecies comparison) than within a species (intraspecies comparison). Despite this tendency, Xiao et al. could also identify several examples of conserved patterns of histone marking across species. By comparing histone mark co-occurrences between species and by fitting histone-based regression models to gene expression levels, Xiao et al. could provide evidence for the regulatory signal

[1] Long et al. [146] use the neutral term of non-methylated islands instead of CGIs, because sequence-based definitions of CGIs may not be directly applicable to all the examined vertebrates.

[2] In the context of transcription factors, “turnover” can be defined as “nucleotide substitution that leads to a transcription factor binding site (TFBS) motif ’moving’ along the DNA sequence in a locus, or ’transforming’ into a different TFBS motif” [149]. Specifically in the context of the work by Schmidt et al., the consequence of a turnover event can be that, e.g., no TF binding motif conservation is detectable between two species, but the local regulatory state may be functionally conserved because nearby (newly formed) TF binding sites could potentially compensate for the lost binding site. Here, it is important to keep in mind that Schmidt et al. examined binding events, and not just binding motifs, which facilitates a more functional interpretation of their results.

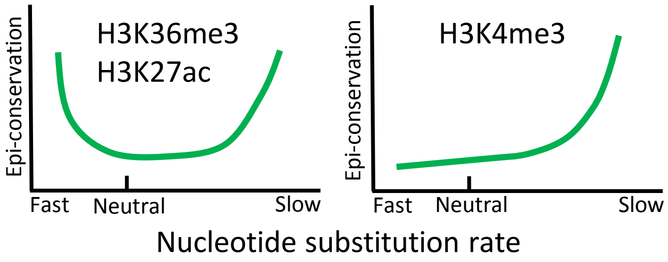


Figure 5.1: Patterns of epigenome conservation: partial reproduction of Figure 3B in Xiao et al. [261]. Schematic illustration of observed patterns of epigenome conservation for the three histone marks relevant in this chapter. Genomic sequence conservation is depicted along the x-axis and (scale-free) epigenome conservation ranging from "low" to "high" is depicted along the y-axis.

represented by certain histone marks seeming to be evolutionarily conserved. A closer inspection of epigenome and genome sequence conservation revealed several patterns of epigenome conservation (Figure 5.1). One of these patterns shows a U-shaped correlation between genome and epigenome conservation, which motivated Xiao et al. to hypothesize that the epigenome could buffer against otherwise deleterious changes on the sequence level (Figure 5.1, left panel). This U-shaped correlation was detected for two of the histone marks relevant in this chapter (H3K27ac, H3K36me3). For the third histone mark used in this chapter (H3K4me3), Xiao et al. observed a different pattern of weak to moderate epigenome conservation in most sequence contexts, but showing a spike in epigenome conservation for highly conserved sequences. The results presented by Xiao et al. seem to suggest that epigenome conservation is not a general consequence of genomic conservation; it is, however, also supported by their evidence that taking sequence conservation as a proxy for epigenome conservation is not an unreasonable starting point to explore epigenome conservation across species.

The conclusions by Xiao et al. [261] seem to be supported by a more recent study of cross-species epigenome evolution. Villar et al. [249] profiled the two histone marks H3K4me3 and H3K27ac to investigate epigenetic changes in enhancer and promoter elements in the liver of 20 mammals. Villar et al. showed that epigenome conservation seems to depend on the regulatory context, similar to what was reported by Xiao et al. [261]. The central result by Villar et al. was stated as follows: epigenome evolution is fast at enhancer elements (H3K27ac) and slow at promoter elements (H3K4me3 and H3K27ac). Restricting the analysis to the 10 species with the highest quality genome assemblies, with human used as reference species, enabled the identification of approximately 300 highly conserved enhancers (of roughly 29,000 in total), and of around 1,800 highly conserved promoters (of roughly 12,000 in total). For these highly conserved regulatory elements, Villar et al. detected a moderate selective constraint on the respective genomic sequences, i.e., these sequences contained fewer nucleotide substitutions than expected under neutral evolution. Nevertheless, Villar et al. characterized the enhancer repertoire in each species as mostly recently evolved and thus largely species-specific.

In summary, the *status quo* of cross-species epigenomics research that looks beyond canonical model animals suggests that there is detectable epigenome conservation on the (histone) chromatin level, with a more or less pronounced bias toward sequence-conserved regions depending on the histone mark. Furthermore, it seems that species-specific changes in the (epigenetic) regulatory landscape are often driven by turnover of TF binding events, and gain or loss of regulatory elements (of course, these events are likely to reciprocally affect each other, or to be just two sides of the same coin). It thus stands to reason that bioinformatics could enable cross-species analyses by harvesting existing epigenome resources, and, consequently, avoid or delay the laborious task of mapping the epigenome in non-model species.

The objective of the work presented in this chapter was to explore and evaluate cross-species extrapolation of epigenome data, and epigenome-based inference of gene expression in a range of target species, based only on existing reference epigenome maps for human and mouse. To that end, we established computational epigenome transfer and prediction of gene expression (Figure 5.2) for twelve mammalian and one avian species (Figure C.1). We transferred epigenome data from our reference species (human or mouse) to the target species using whole-genome alignments. We used the transferred epigenome data to predict tissue-specific gene expression in the target species using machine learning models trained and cross validated on data from the reference species. To validate our approach, we compared our predictions with measured tissue-specific expression profiles of the target species, confirming that such cross-species predictions can be useful and informative.

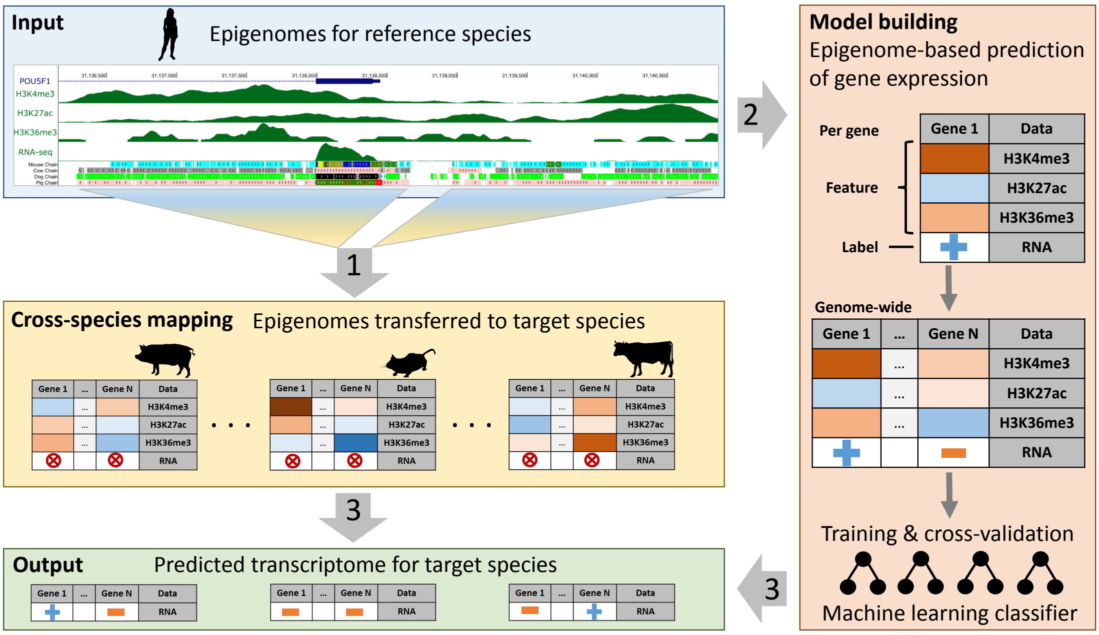

Input
Epigenomes for reference species
POU5F1
H3K4me3
H3K27ac
H3K36me3
RNA-seq
Mouse Chain
Cow Chain
Dog Chain
Pig Chain
1
Cross-species mapping
Epigenomes transferred to target species
Gene 1
...
Gene N
Data
H3K4me3
H3K27ac
H3K36me3
RNA
2
Model building
Epigenome-based prediction of gene expression
Per gene
Feature
Label
Genome-wide
Training & cross-validation
Machine learning classifier
3
3
Output
Predicted transcriptome for target species

Figure 5.2: Conceptual outline of the cross-species pipeline: (1) top to middle: epigenome data from the reference species are transferred to the target species using pairwise whole-genome alignments. (2) Top to right: epigenome data in the reference species are used to train machine learning classifiers that predict gene expression status (blue cross: on/high; orange minus: off/low; red “no” sign: unlabeled data). The strength of the epigenetic signal in gene loci (here illustrated ranging from blue/low to red/high) is used as model feature. (3) Bottom: prediction of gene expression in the target species using the transferred epigenome data (yellow box) and the trained machine learning classifier (red box).

5.2 Materials and Methods

5.2.1 Included species, genome assemblies, and gene models

We included 12 mammalian species (human [hg19], rhesus [rheMac2], mouse [mm9], rat [rn5], rabbit [oryCun2], pig [susScr2], cow [bosTau7], sheep [oviAri3], horse [equCab2], dog [canFam3], cat [felCat5], opossum [monDom5]), and one avian species (chicken [galGal3]), based on the following selection criteria: (i) complete genome assemblies and whole-genome alignments with at least one of the two reference species (human, mouse) were available from the UCSC Genome Browser [271]; (ii) gene models for the relevant assemblies were available from one of three sources (GENCODE [96]: human v19, mouse vM1; The Bovine Genome Database [69]: cow Ensembl v75; UCSC Genome Browser tables ensGene, ensemblSource and ensemblToGeneName: all other species); (iii) epigenome profiles including histone H3K4me3, H3K27ac, and H3K36me3 as well as transcriptome data were available for defined tissues/cell types in the reference species (Appendix C.1, Table S1); and (iv) transcriptome data were available for at least some of these tissues/cell types in the target species. An overview of the evolutionary relationships for all species included in this study is provided as a phylogenetic tree generated using the TimeTree service [99] C.1. To alleviate the effect of differences in annotation quality, all gene models were reduced to protein-coding transcripts/genes. Additionally, only transcripts tagged as "Consensus CDS (CCDS)" were selected in the GENCODE annotations. All analyses were restricted to genes located on the autosomes. Promoter regions were defined as 1.5 kilobase windows around the TSS (-1,000 bp to +500 bp), and genes with a gene body length of less than 750 bp were discarded.

5.2.2 Whole-genome alignments, gene orthologs, and evolutionary conservation

Whole-genome alignments between the reference species (human, mouse) and target species were downloaded from the UCSC Genome Browser in the form of chain files [122]. Following the instructions in the UCSC Genome Wiki[3], the downloaded chain files were processed to derive so-called reciprocal best chains, which represent a reduced (high-quality) alignment relative to the complete alignment information stored in the chain files. The reciprocal best chains were further processed using CrossMap [269] and custom scripts to build pairwise symmetric alignment blocks. Genes with less than 100 aligned bases in their promoter and in their body were considered weakly aligned. Information on gene orthologs was downloaded from OrthoDB [133], and lists of 1-to-1 orthologs for each pair of species and, separately, for all 13 species combined were extracted from the annotation data using custom scripts.

5.2.3 Epigenome and transcriptome data preprocessing

Publicly available reference epigenomes for the reference species (human, mouse) were obtained from ENCODE, DEEP, and BLUEPRINT (Appendix C.1, Table S1). The resulting dataset included three histone marks (H3K4me3, H3K27ac, H3K36me3), three cell types (embryonic stem cells, naïve CD4+ T cells and hepatocytes), and a total of nine epigenome profiles for human. For mouse, the dataset included three histone marks (H3K4me3, H3K27ac, H3K36me3), five cell/tissue types (embryonic stem cells, naïve CD4+ T cells, whole liver, kidney and heart), and a total of 17 epigenome profiles. Tissue-specific transcriptome profiles were obtained from ENCODE, DEEP, and public repositories (Sequence Read Archive (SRA)/European Nucleotide Archive (ENA)). Where available, epigenome profiles were downloaded in the form of histone signal tracks (bigWig format; see Chapter 2, Section 2.4 for details). For the BLUEPRINT mouse data, which were not available in preprocessed form, reads were mapped using bowtie2 v2.3.3.1 with the preset "`-sensitive`", and signal tracks were generated using bamCoverage v2.5.3 from the deepTools software suite [199, 200]. To prepare the epigenome profiles for the analysis, biological replicates from the same laboratory were merged by taking the mean. All resulting epigenome signal tracks were quantile normalized per project and clipped at the 99.95 percentile to alleviate the effect of outliers. For validation purposes, the liver epigenome data by Villar et al. [249] (ENA study accession PRJEB6906) were downloaded and histone signal tracks for H3K4me3 and H3K27ac were generated following the same procedure as for the BLUEPRINT mouse reference epigenomes.

[3] genomewiki.ucsc.edu/index.php/HowTo:_Syntenic_Net_or_Reciprocal_Best

To evaluate the predictions of our machine learning models (Section 5.2.4) with experimentally determined gene expression levels, we obtained transcriptome data from various public sources: embryonic stem cells for human and mouse [231, 241]; CD4+ T cells for human and mouse [46, 60]; hepatocytes for human [48]; whole-organ samples of liver, kidney and heart for all species except human [81, 112, 162]; and a blood sample for opossum (Appendix C.1, Table S1). Transcriptome data were processed with Salmon v0.8.2 using the following parameters: "`-forgettingFactor 0.8 -useVBOpt -seqBias -gcBias -geneMap`", aggregating transcript-level abundance estimates into gene-level estimates. Finally, the gene expression values were subjected to quantile normalization, and the normalized transcript per million (TPM) values were used for further analysis.

5.2.4 Epigenome-based prediction of gene expression

The formal description of the machine learning procedure follows Hastie et al. [97] in argument and notation if not stated otherwise

All prediction models were implemented in Python3, using libraries from the SciPy ecosystem for scientific computing [160, 166, 177, 183]. Histone signal tracks were masked to exclude non-conserved regions according to pairwise genome alignments between the reference and target species. Prediction attributes were derived from these masked signal tracks by averaging the signal across each gene promoter (H3K4me3, H3K27ac) and gene body (H3K36me3). The machine learning part of our pipeline was realized by training gradient boosting classifiers from the scikit-learn library [183]. The general idea of boosting is to combine many weak learners into a more capable and robust ensemble method; here, a weak learner is any (simple) base model that has a performance only slightly better than random guessing. The most popular choice for selecting the base model are decision trees (cf. Hastie et al. [97], Section 10.7), which is also the base model used in the scikit-learn implementation. The fitting of a gradient tree boosted model is a forward stagewise process, i.e., an iterative procedure that adds one base model at a time without changing the already existing models in the ensemble. Formally, given a training dataset of N samples, with feature matrix X and class labels Y, this can be expressed as follows using decision tree models T as base learners:

$$T(x;\Theta) = \sum_{j=1}^{J} \gamma_j \mathbb{1}\left[x \in R_j\right] \tag{5.1}$$

where $\Theta = \{R_j, \gamma_j\}_1^J$. A tree T partitions the feature space into J disjoint regions R_j that are represented as terminal nodes of the tree. In the case of $J = 2$, such a tree model is called a "decision stump" that makes only a single split. For each region, there is a constant γ_j and the predictive rule is

$$x \in R_j \Rightarrow f(x) = \gamma_j \tag{5.2}$$

In classification scenarios, an intuitive way to choose γ_j would be, e.g., to use the label of the modal class in region R_j. Details on how to find the R_j, i.e., a "good" partitioning of the space into disjoint regions are omitted here (see Hastie et al. [97], Section 9.2). The boosted tree model f_M is then a sum over a total of M trees T:

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m) \tag{5.3}$$

where each step of the forward stagewise fitting amounts to solving

$$\hat{\Theta} = \arg\min_{\Theta_m} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \Theta_m)) \tag{5.4}$$

The loss function L is the deviance/cross-entropy in our implementation (cf. Hastie et al. [97], formula 9.17)

$$K \text{ classes} : -\sum_{k=1}^{K} p_k \cdot log(p_k) \tag{5.5}$$

$$K = 2 : -p \cdot log(p) - (1-p) \cdot log(1-p) \tag{5.6}$$

where p and $1-p$ are, in our setting of binary classification, the class proportions in a tree node representing a region R_j. An important property of gradient boosting is that each subsequent stage in the fitting process is guided by previous “mistakes”. This is realized by fitting the m-th model f_m to the negative gradient of the loss function evaluated at f_{m-1}. Let g_{im} be the gradient for sample i in iteration m and squared error used to measure the closeness of fit of the m-th model to the negative gradient, this leads to[4]

$$g_{im} = [\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f=f_{m-1}} \tag{5.7}$$

$$\hat{\Theta}_m = \arg\min_{\Theta} \sum_{i=1}^{N} (-g_{im} - T(x_i; \Theta))^2 \tag{5.8}$$

A pseudocode version of the complete fitting procedure can be found in Hastie et al. [97] (Algorithm 10.3), of which we just repeat the update rule completing iteration m:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{im} \mathbb{1}\left[x \in R_j\right] \tag{5.9}$$

A common regularization parameter used for gradient tree boosting is the so-called learning rate ν. Introducing this parameter into the algorithm changes the update rule 5.9 to

$$f_m(x) = f_{m-1}(x) + \nu \cdot \sum_{j=1}^{J} \gamma_{im} \mathbb{1}\left[x \in R_j\right] \tag{5.10}$$

[4] Equation 5.8 omits a technical detail related to the (inexact) approximation to the negative gradient. Since this chapter only describes an application of the gradient tree boosting procedure, and more elaborate technical considerations are beyond the scope of this thesis, this minor simplification from Hastie et al. [97] seems to be a reasonable compromise.

For values smaller than one, ν limits the contribution of each individual base learner. Limiting the contribution of the base learners leads to a larger value of M to achieve the same error rate during model training. Despite this trade-off between ν and M, empirical evidence suggests that it is usually beneficial to tune ν as an additional hyperparameter during cross validation (see Hastie et al. [97], Section 10.12.1). We followed this advice in the model building part of our pipeline.
Gradient boosting classifiers were trained using histone signal intensities as prediction attributes and the gene expression status (on/high: TPM ≥ 1; off/low: TPM < 1) as target variable (this binary thresholding strategy was motivated by previous studies [34, 58, 227]). Each training dataset was randomly subsampled to balance class frequencies, and model hyperparameters J, M and ν were tuned using five-fold cross validation on this subsampled training dataset. The best model according to the results of the cross validation was refit using the full set of training data. For the evaluation of the trained classifiers, the model performance was measured in terms of accuracy as defined in Chapter 4, Section 4.2.5. Additionally, the terms sensitivity and specificity are used in this chapter. Sensitivity is the same as recall (cf. Chapter 4, Section 4.2.5), and specificity is the true negative rate defined as the number of true negative samples divided by the sum over all true negative and false positive samples.

5.2.5 Genomic region enrichment analysis

Region sets were analyzed for significant enrichment using the LOLA software [225]. For the human genome, the LOLA Core region database was used. In addition, we created a custom region set for both the human and mouse genome, comprising various sets of transcription factor binding sites as well as histone modification peaks from the DeepBlue repository [5]. For each LOLA analysis, we filtered the results and retained enriched region sets if the support (i.e., number of regions covered) was at least five and the multiple-testing corrected statistical significance (q-value [232]) was below 0.05. We manually selected 10–15 entries from the top ranking region sets for visualization and provide the full list of LOLA enrichments in the online supplement (Appendix C.1, Tables S2 and S3).

5.2.6 Gene age annotation

To evaluate the evolutionary age of each gene, we obtained gene age annotations for the following species from a recent publication [142]: chicken, cow, dog, human, mouse, opossum, rat, and rhesus macaque. UniProt identifiers were mapped to Ensembl gene identifiers using a web–based service [195]. Since the identifiers were mapped with varying success depending on the species, the results presented below focus on the the three species where at least 80% of all gene identifiers could be mapped (human 97%; mouse 83%; rhesus 81%).

5.2.7 Study replication

The code for replicating all results of this chapter is publicly available at doi.org/10.17617/1.69. All data preprocessing and analysis pipelines have been implemented in Python/Ruffus [90]. All figures except for Figure 5.2 can be recreated using the respective Jupyter Notebooks[5] available in the aforementioned repository.

5.3 Results

5.3.1 Cross-species transfer of reference epigenome data using whole-genome alignments

Our bioinformatics pipeline uses epigenome profiles for the reference species as input and exploits whole-genome alignments to transfer these data to conserved genomic regions in the target species. To that end, we prepared pairwise symmetric whole-genome alignments for each pair of reference and target species (Section 5.2.2), and we used these alignments to transfer histone signal intensities. As motivated in Section 5.1, our approach is based on the hypothesis that sequence conservation is an indicator of regulatory conservation within a genomic region, and that tissue-specific patterns of epigenome regulation are frequently maintained across species in sequence-conserved regions. We tested and confirmed this hypothesis by using the transferred epigenome data to predict tissue-specific gene expression in the target species — as described further below.

Using alignment-based epigenome transfer, we produced genome-wide, cell-type specific epigenome profiles for each target species. The non-zero part of the transferred histone signal covers a substantial fraction of the overall aligned bases in the target species (ranging from human and mouse to opossum and chicken), based on the two reference species human (Figure 5.3) and mouse (Figure C.2). The distribution of the transferred histone signal follows patterns similar to those of the measured histone signal in the reference species: H3K4me3 signal covers the smallest number of bases, consistent with this mark's prevalent occurrence in gene promoters. The more widespread occurrence of H3K27ac and H3K36me3 is similarly consistent with the focus of H3K27ac on a broader set of regulatory regions (including enhancer elements) compared to H3K4me3, and H3K36me3 covers the gene body of actively transcribed genes (Chapter 2, Section 2.1.1.4).

Further analysis of average histone signal strength in gene promoters, gene bodies, and other genomic regions showed that, after cross-species transfer, the histone signal strength in the target species resembles the distribution in the reference species (Figure C.3). The transferred signal is generally strongest in gene bodies and in gene promoters, while it is weak outside of protein-coding genes. Since we only used the histone signal in gene promoters and in gene bodies for predictive modeling of gene expression, the low histone signal outside of these two region types is of no concern in the prediction analysis described below.

As a validation for the cross-species transfer of epigenome data, we examined whether

[5] http://jupyter.org

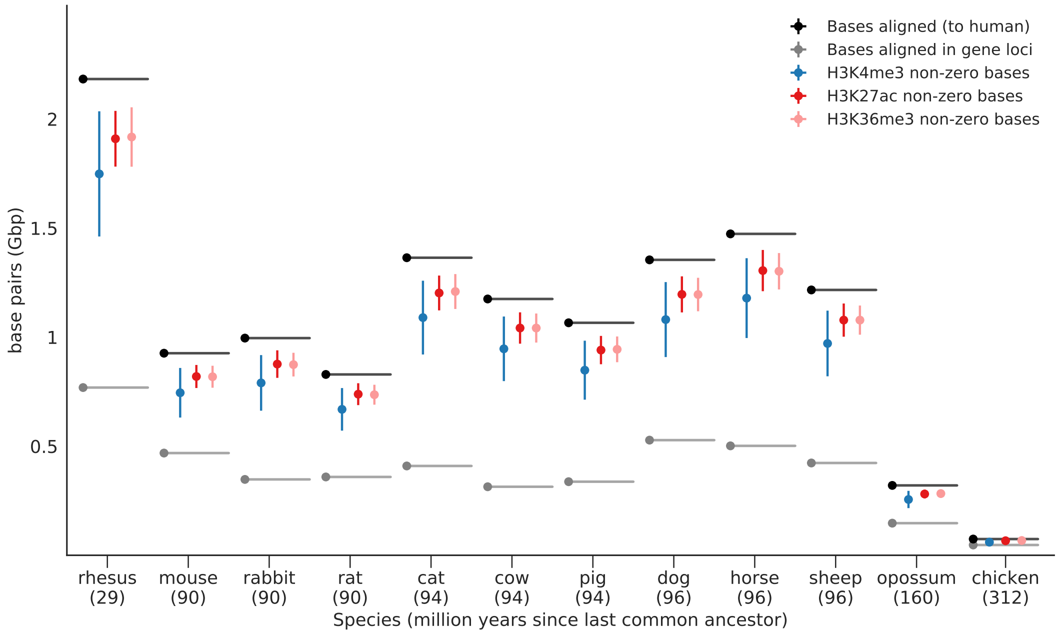


Figure 5.3: Coverage of transferred epigenome profiles: colored points indicate the average number of base pairs with non-zero signal after cross-species transfer for each histone mark (error bars denote +/- one standard deviation around the mean) between human as the reference species and twelve target species. The number of aligned bases genome-wide (black bars), and at gene loci (gene promoters and gene bodies combined, grey bars) are shown for comparison. Target species are sorted by increasing evolutionary distance to the human reference (x-axis, million years since last common ancestor indicated in parentheses).

the transferred epigenetic signals in gene promoters (for H3K4me3 and H3K27ac) and gene bodies (for H3K36me3) retain detectable cell-type specificity. To that end, we focused on the comparison between human and mouse, where we have reference epigenome data for both species and most samples, and we found that correlations between transferred and measured epigenomes are usually highest when the cell type is matched (Figure 5.4). For example, the comparison between transferred and measured epigenomes in mouse results in a Pearson correlation of 0.75 for H3K4me3, 0.61 for H3K27ac, and 0.71 for H3K36me3 in CD4+ T cells.

Although in general epigenome data is scarce for non-model organisms, we could use the data generated by Villar et al. [249] to corroborate the correlations observed above in one tissue (liver) for at least two (H3K4me3, H3K27ac) of the three histone marks in several target species. We computed the histone signal correlation in gene promoters for both marks comparing transferred to measured epigenome data. Additionally, for the two reference species human and mouse, we also correlated the data by Villar et al. to our DEEP and ENCODE reference epigenomes (Figure 5.5). The data by Villar et al. show strong agreement with the DEEP and ENCODE reference epigenomes in human and mouse, with the lowest median correlation at 0.89 for the human H3K27ac signal. The correlations are slightly lower when comparing the data measured by Villar et al. to the transferred epigenome data in the respective target species. However, the median correlation is still high at approximately 0.7 or higher in most cases. Transferred epigenomes having pig as target species show the lowest correlations of around 0.6 to 0.65 for both histone marks and for both reference species.

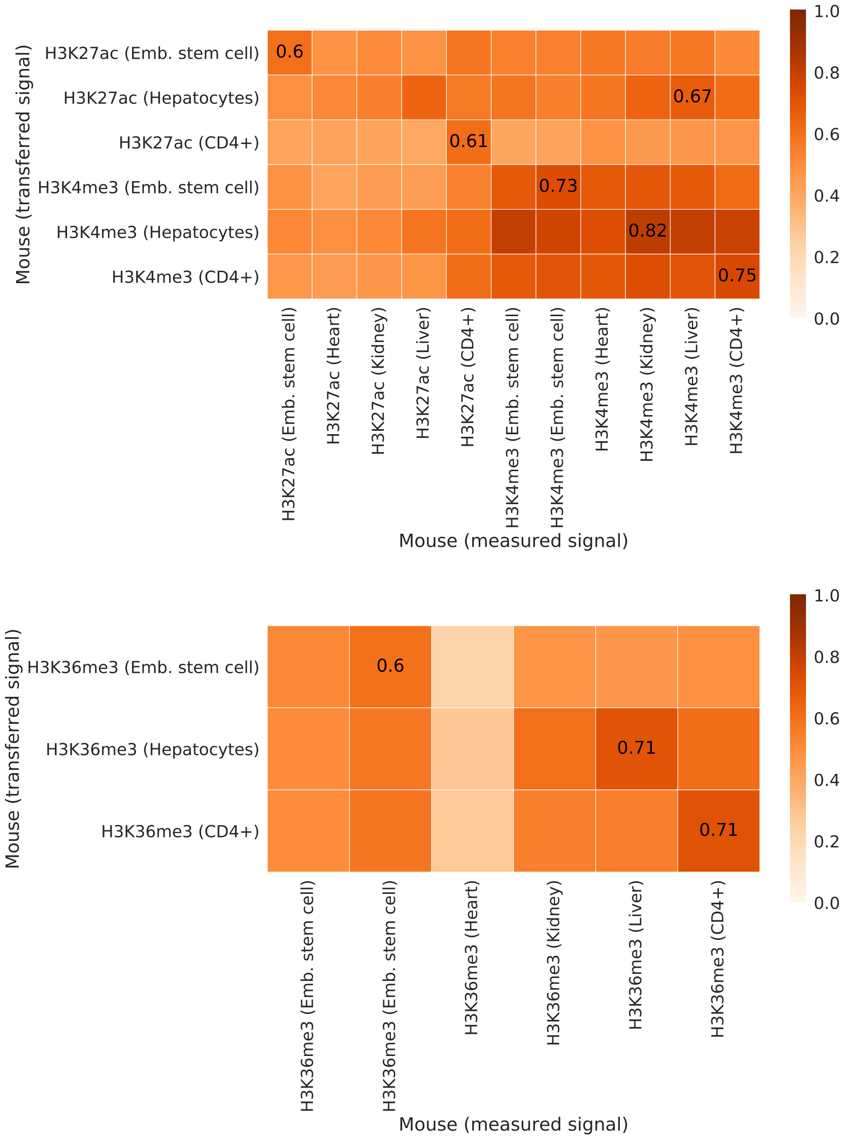


Figure 5.4: Tissue-specific correlation of measured and transferred epigenomes: heatmaps show pairwise Pearson correlations of measured and transferred histone signals for H3K4me3 and H3K27ac at gene promoters (top panel) and for H3K36me3 in gene bodies (bottom panel), using human as reference species and mouse as target species.

To substantiate these findings, we performed region set enrichment using the LOLA software [225] on the top 5% gene promoters with the highest average transferred signal for H3K4me3, and on the top 5% gene bodies with highest average signal for H3K36me3. We consistently observed cell-type specific enrichment (Appendix C.1, Table S2), as illustrated by the LOLA enrichment for CD4+ T cells using mouse as reference and human as target species (Figure 5.6A/B) and by the LOLA enrichment for liver/hepatocytes using human as reference and mouse as target species (Figure 5.6C/D). In both cases, the LOLA results show the expected enrichment for tissue-specific histone marks associated with active promoters (H3K4me3) and actively transcribed genes

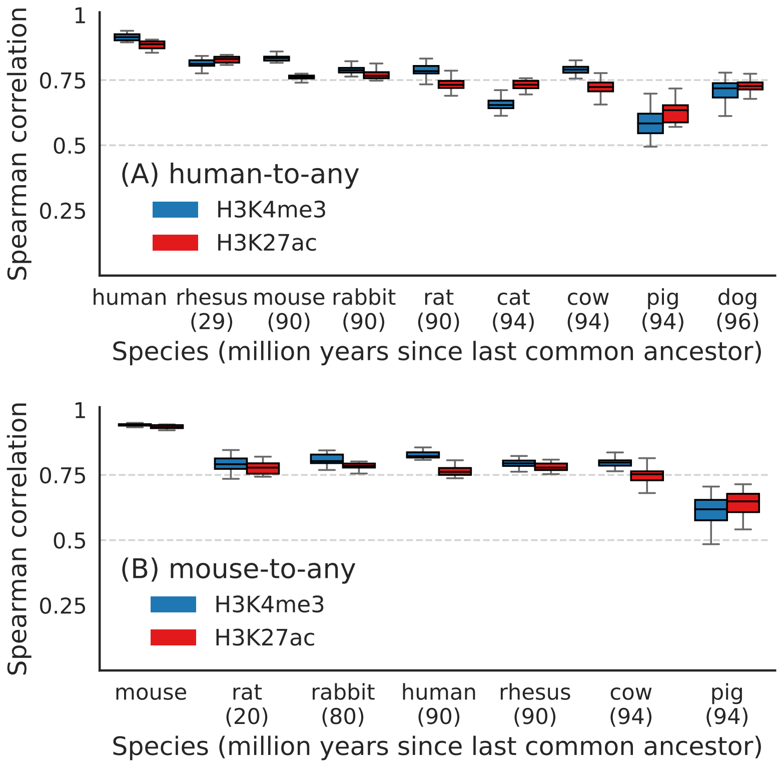


Figure 5.5: Correlation of measured and transferred liver epigenome data: box plots show Spearman correlation of measured and transferred histone signals for H3K4me3 and H3K27ac at gene promoters using human (A) and mouse (B) as a reference for the cross-species epigenome transfer. Measured liver epigenome data were obtained from the study by Villar et al. [249]. In the case of human-to-human and mouse-to-mouse, the validation data by Villar et al. were correlated to the DEEP and ENCODE reference epigenome data used in our study. Box plots depict correlation values together for all individual chromosome in the target species.

(H3K36me3), respectively. The annotated cell type of the enriched region sets is the same as (or closely related to) the cell type of the transferred epigenomes, supporting that our cross-species epigenome transfer method retained the characteristic cell-type specificity of epigenome data.

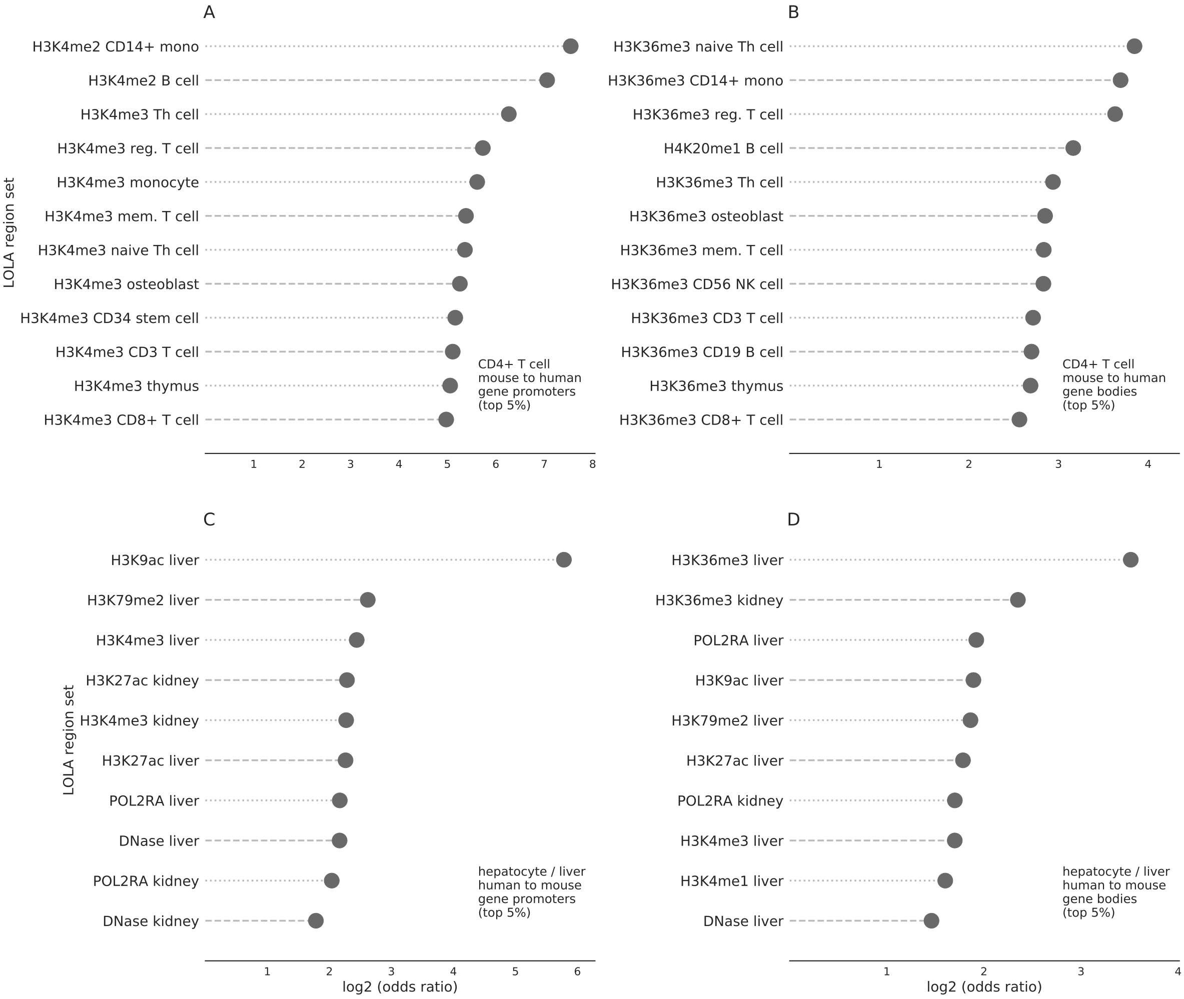


Figure 5.6: Genomic region enrichment analysis for transferred epigenomes: selected results of LOLA analyses for gene promoters (panel A and C) and gene bodies (panel B and D) that were ranked among the top 5% based on transferred epigenome signal intensities for H3K4me3 (gene promoters) and H3K36me3 (gene bodies). Top panels show region set enrichment for CD4+ T cell epigenomes transferred from mouse to human (panel A and B); and bottom panels show region set enrichment for hepatocyte/liver epigenomes transferred from human to mouse (panel C and D). Effect size (log2 of the odds ratio) is indicated on the x-axis. The false discovery rate estimate (q-value [232]) is smaller than 10^{-12} in all cases.

5.3.2 Prediction of gene expression using epigenome data transferred across species

To assess the biological information carried by the transferred epigenomes in a larger number of species (namely in those species for which hardly any epigenome data are available for validation), we tested whether we could predict gene expression in the target species based on the transferred epigenomes. It is well-established that gene expression can be predicted from epigenome data [25, 115, 227]. We would expect to observe better-than-random accuracies when predicting gene expression using transferred epigenome profiles if these profiles indeed captured relevant regulatory biology. Moreover, to make the validation more stringent, we can exploit the cell-type specific character of epigenome data and try to predict cell-type specific patterns of gene expression.

We used the uniformly processed transcriptome dataset described in Section 5.2.3 as our experimental reference dataset against which we evaluated the epigenome-based predictions. In this transcriptome dataset, we observed consistent clustering by cell type rather than by species, both for 1-to-1 gene orthologs [133] between each pair of species (Figure 5.7) and for those genes that were conserved across all 13 species (Figure C.4). It should be noted that there is some controversy concerning the conservation of gene expression in orthologous genes, and different studies reported either clustering by tissue or clustering by species [31, 144, 270].

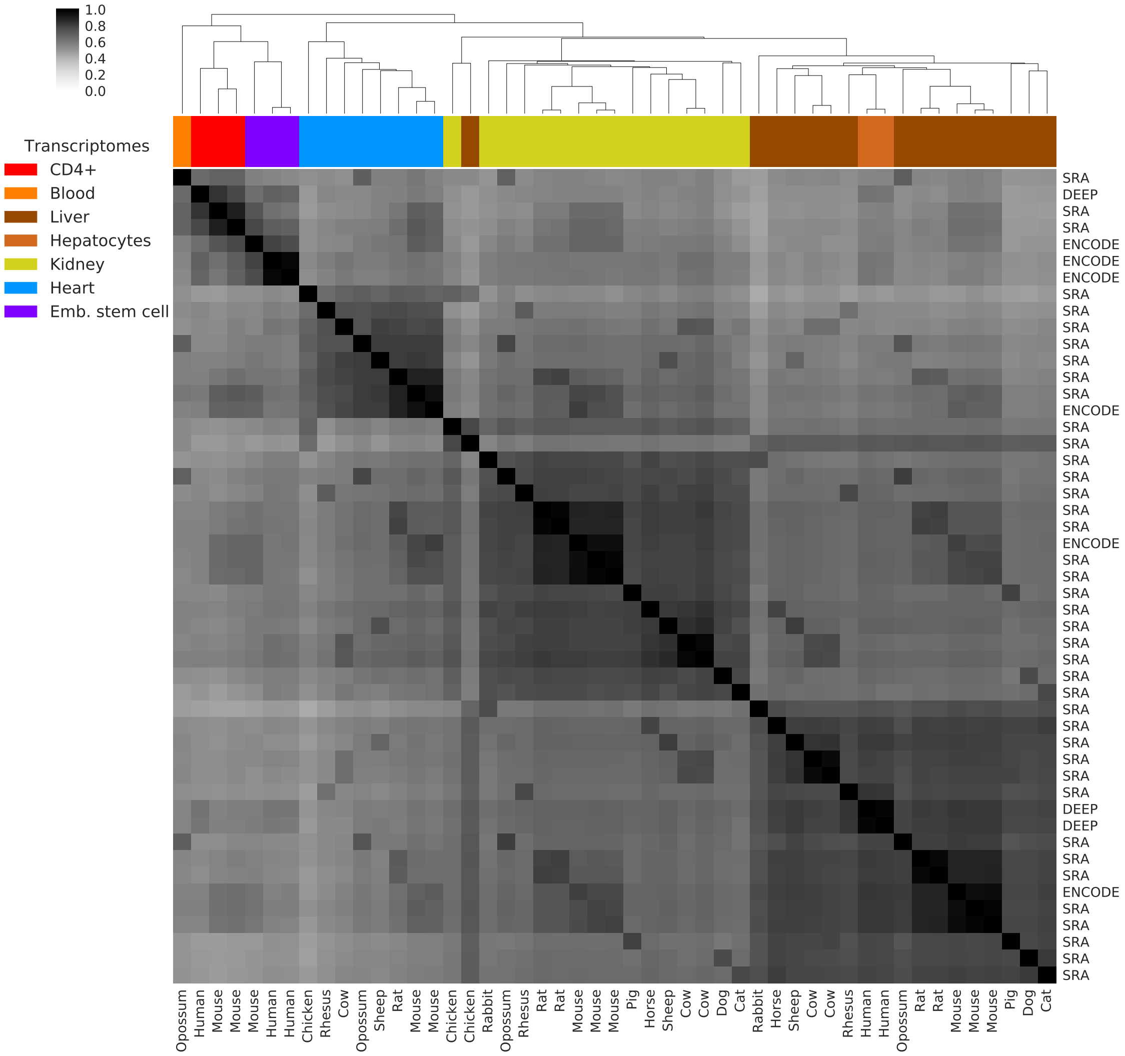


Figure 5.7: Transcriptome data cluster by tissue of origin: hierarchical clustering of pairwise Pearson correlation coefficients between transcriptome profiles of 13 species used for gene expression prediction and model validation. Correlations were calculated across all 1-to-1 orthologous genes for each species pair. Color bars at the top indicate tissue of origin. Columns are labeled with species of origin, and row labels indicate data source (project name or "SRA" for data downloaded from SRA/ENA).

We implemented a machine learning approach to predict gene expression profiles in the target species based on epigenome data in the reference species (see Figure 5.2 for an overview). Using only data from the reference species, binary classifiers were trained to predict gene expression as either on/high or off/low (Section 5.2.4). The prediction attributes were obtained by averaging histone signals for H3K4me3/H3K27ac in gene promoters and for H3K36me3 within the gene body, restricted to those subregions that were covered by the cross-species alignment used for epigenome transfer. A threshold of one transcript per million (1 TPM; see Section 5.2.4) was applied to label genes as on/high or off/low. We observed consistently high prediction performance for the two reference species, with cross-validated, test-set only, area under the receiver operating characteristic curve (AUC-ROC; in the following: AUC) values of 0.89 (human to mouse) and 0.90 (mouse to human), resulting in a sensitivity of 78% (human) and 76% (mouse) at a specificity of 90% (Figure 5.8).

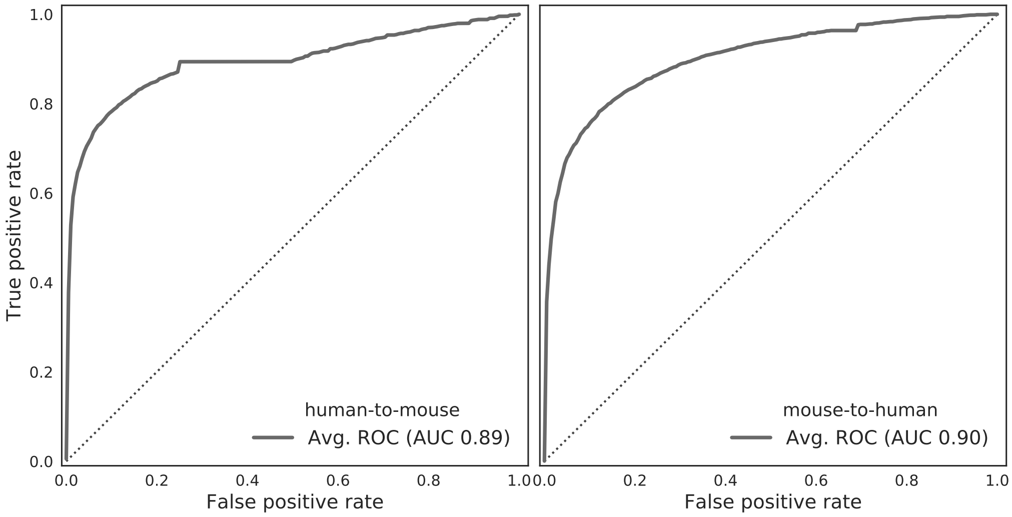


Figure 5.8: Gene expression prediction performance for human/mouse: ROC curves show the cross-validated test-set performance predicting gene expression status in mouse based on transferred epigenome data from human (panel A) and vice versa (panel B), averaged across all included samples and cell types in both the reference and the target species. The dashed diagonal line represents the expected performance of a random classifier (AUC = 0.5).

Having validated the epigenome-based classifiers in the reference species, we applied the classifiers to all target species, using the transferred epigenome data as input. We predicted gene expression in each target species independent of cell type, averaging across transferred epigenome profiles and transcriptomes in each target species. We observed high prediction accuracies for all target species, with AUC values ranging from 0.87 to 0.81 when using human as reference (Figure 5.9, left), and from 0.89 to 0.83 when using mouse as reference (Figure 5.9, right). The average sensitivity is 67% (human reference) and 73% (mouse reference) at a specificity of 90%.

As an additional evaluation, we tested if our gene expression predictions are cell-type specific, i.e., they should show the highest accuracy for the matched cell type between reference and target species. When comparing AUC values of tissue-matched prediction to AUC values obtained by cross-tissue prediction, the tissue-specific predictions consistently outperform the tissue-agnostic

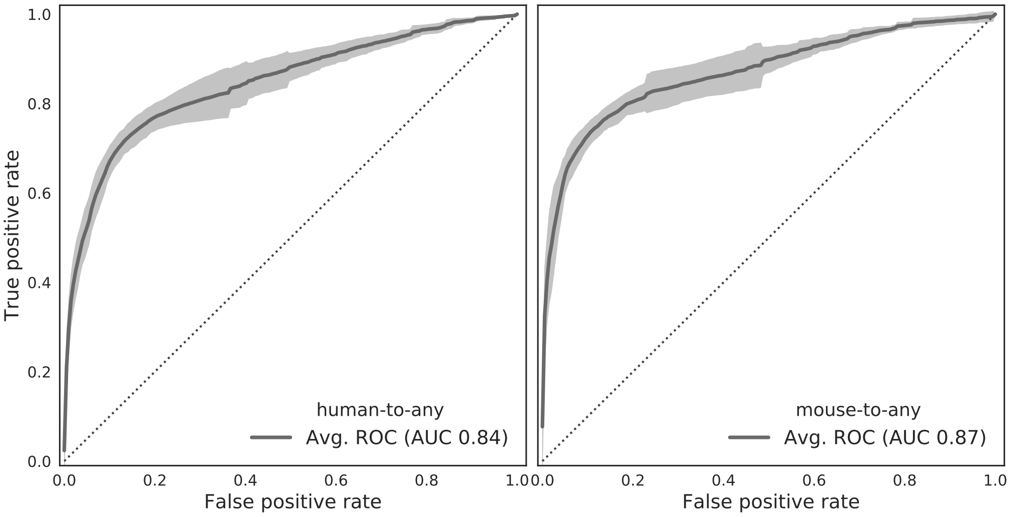


Figure 5.9: Gene expression prediction performance for all species: ROC curves show the cross-validated test-set performance predicting gene expression status based on transferred epigenomes from human (left panel) or mouse (right panel) as reference species and the remaining twelve species as target species. Epigenomes and transcriptomes were averaged across all included samples and cell types in both the reference and the target species. Shaded areas represents +/- 1 standard deviation around the mean ROC curve. The dashed diagonal line represents the expected performance of a random classifier (AUC = 0.5).

predictions across all investigated target species and for both human (Figure 5.10, top panel) and mouse (Figure 5.10, bottom panel) as our reference species. Aggregating AUC values across species, the difference between the mean AUC values for the tissue-matched tests and the cross-tissue controls is highly statistically significant (one-sided Mann-Whitney-U, p-value $< 10^{-9}$).

5.3.3 Comparison of epigenome-based and orthology-based prediction of tissue-specific expression

To benchmark our epigenome-based predictions of gene expression, we compared their performance to that of an alternative (and complementary) approach that is based on gene orthology. Specifically, for all 1-to-1 gene orthologs between each pair of species, the orthology-based method predicts a gene to be expressed in a certain cell type of the target species if — and only if — it is expressed in the corresponding cell type of the reference species. This method can predict expression only for genes that have an annotated 1-to-1 ortholog in the reference species, which limits its scope and applicability to a subset of genes (Figure C.5).

For a systematic comparison between the epigenome-based and orthology-based methods, we evaluated their performance on various subsets of genes. These subsets were constructed using the following approach: we ordered all genes in the target species by increasing levels of DNA sequence conservation in the gene body, and we used all the genes above a certain threshold as our evaluation set. We then plotted the performance gain of the epigenome-based method (calculated

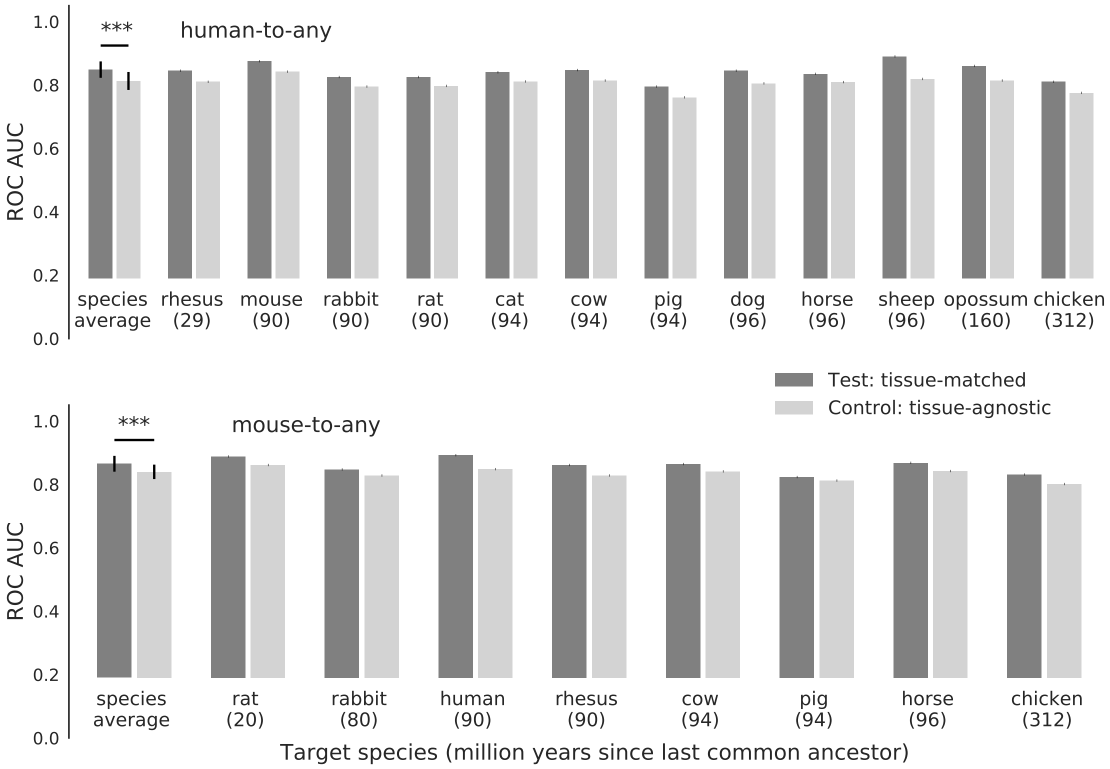


Figure 5.10: Tissue-specific gene expression prediction performance: AUC values (calculated as in Figure 5.9) comparing the performance of gene expression prediction based on tissue-specific (dark grey) and on average (light grey) transferred epigenomes, using human (top panel) or mouse (bottom panel) as reference species and the remaining twelve species as target species. The first bar shows the aggregation of mean AUC values across all species. The difference in mean AUC for these aggregated values is statistically significant for both reference species (one-sided Mann-Whitney-U, “***” indicates significance at $p < 10^{-9}$).

as the surplus of correct predictions over the orthology-based method across all target species) for human and mouse as our reference. In this analysis, the epigenome-based method results in approximately 20% more correct predictions than the orthology-based method (Figure 5.11, left panel).

Both methods display stable prediction accuracy over a broad range of gene conservation values (Figure 5.11, right panel). Although the performance gain of epigenome-based prediction is higher for stringent thresholds on gene body conservation (Figure 5.11, left panel), only relatively few genes pass these stringent thresholds, and a more inclusive threshold may increase the scope and utility of our predictions. For example, lowering the threshold on gene body conservation to 10–15% makes it possible for the epigenome-based method to predict the expression status for approximately 90% of all genes in the target species, compared to approximately 80% for the orthology-based method. Given that both methods have similar median accuracies of 75.1% and 79.9% (human reference) and of 71.5% and 78.1% (mouse reference), the epigenome-based method offers an advantage over the orthology-based method by providing a substantially larger number of accurate predictions.

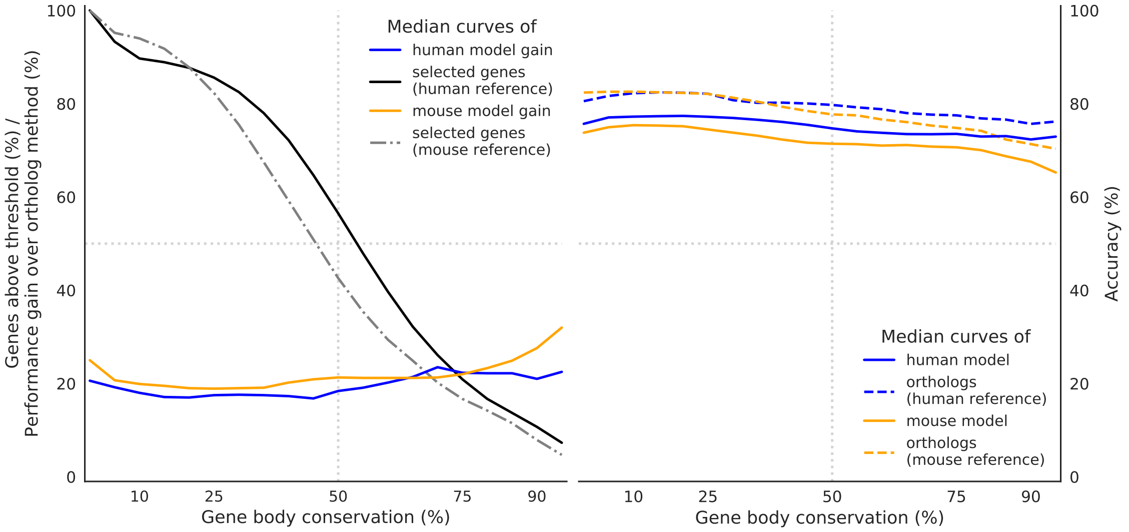


Figure 5.11: Comparison of epigenome-based and orthology-based approach: (left) performance gain of epigenome-based prediction of gene expression over the orthology-based approach using human (blue) or mouse (orange) as reference species, plotted for different thresholds on the gene body conservation (x-axis). The number of selected genes for each threshold is also indicated (black and gray lines). Performance is measured as the number of correct predictions (true positives and true negatives), and the surplus of correct predictions of the epigenome-based prediction models over the orthology-based approach is transformed into a percentage value that is comparable across species. (right) prediction accuracy of epigenome-based (solid lines) and orthology-based (dashed lines) prediction of gene expression (solid lines) and the corresponding orthology-based predictions (dashed lines) using human (blue) or mouse (orange) as reference species, plotted for different thresholds on the gene body conservation (x-axis). All curves (left and right panel) represent median values of all gene expression predictions aggregated per reference species over all target species and over all cell types. Genes in the target species are sorted by increasing gene body conservation along the x-axis.

Finally, we evaluated the tradeoff between sensitivity and specificity in more detail for the epigenome-based method. We interpreted the class probabilities of the classifier as a measure of prediction confidence. We observed that the advantage of the epigenome-based method over the orthology-based method is strongest for lenient thresholds on the class probability of around 0.5, at the cost of a slightly reduced accuracy (Figures 5.12 and C.6). For the most stringent thresholds (> 0.9), the epigenome-based method consistently makes a higher number of correct predictions than the orthology-based method, while the difference in accuracy between both methods approaches zero. Based on the shape of the curves, a relatively stringent threshold of 0.75 is likely to constitute a suitable and broadly applicable tradeoff between sensitivity and specificity.

5.3.4 Limitations of cross-species epigenome data transfer and prediction of gene expression

Despite these promising results using cross-species epigenome transfer and epigenome-based prediction of gene expression, the approach has certain limitations. Most importantly, whole-

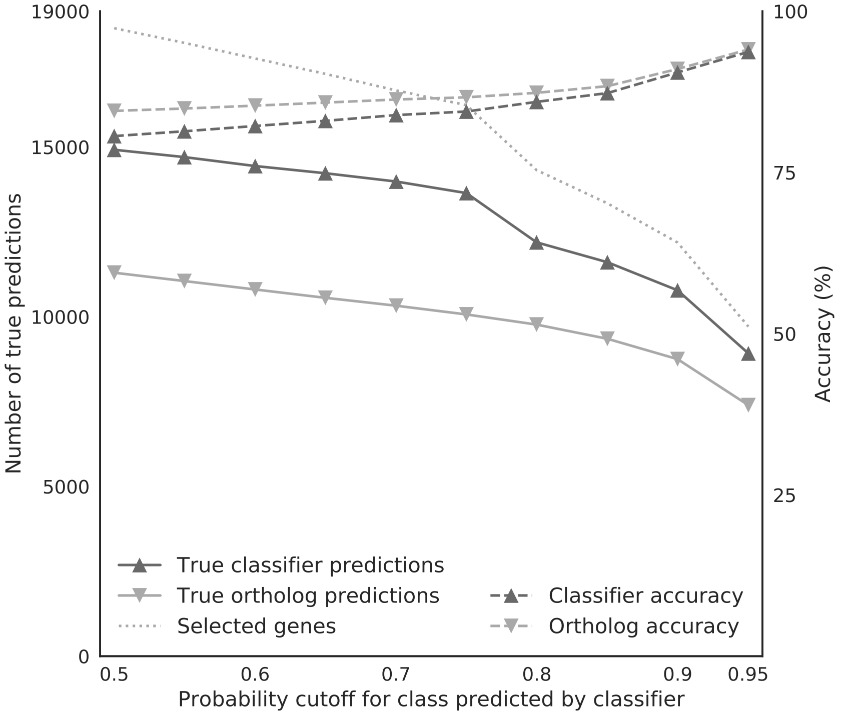


Figure 5.12: Effect of thresholding on classifier class probability estimates: prediction performance of the epigenome-based approach (dark gray lines) and the orthology-based approach (light gray lines), as measured by the number of correct predictions (true positives and true negatives, left y-axis) and percent accuracy (right y-axis) for increasingly stringent cutoffs on the class probability estimates of the epigenome-based approach (x-axis, range from 0.5 to 0.95). The number of genes selected at each threshold is indicated as dotted line. Results are shown for human as reference and mouse as target species.

genome alignments cover only those genomic regions for which there is discernible conservation of the DNA sequence. For example, roughly 1 giga base pairs (Gbp) of DNA sequence are aligned between the human and mouse genome (90 million years since last common ancestor), which corresponds to approximately a third of the human genome. At a threshold of at least 100 conserved base pairs for both the gene promoter and gene body, 92.2% (mouse) and 97.2% (human) of genes can be target for cross-species transfer of reference epigenome data. These values remain high across larger evolutionary time (Figures 5.13 and C.7), for example amounting to 98.8% in the comparison between human and opossum (160 million years) and 90.7% between human and chicken (312 million years).

We investigated the 487 human and 1,435 mouse genes that failed to meet our minimum alignment thresholds and which we therefore cannot predict using the epigenome-based method. Functionally characterization of these gene sets using the enrichR web service [41, 134] identified an enrichment for Gene Ontology categories related to olfactory reception (Appendix C.1, Table S4), which is consistent with large species-specific differences in the repertoire of olfactory receptor genes between human and mouse [173]. We also analyzed the corresponding promoter regions for region set enrichment using the LOLA software [225], and we found an enrichment

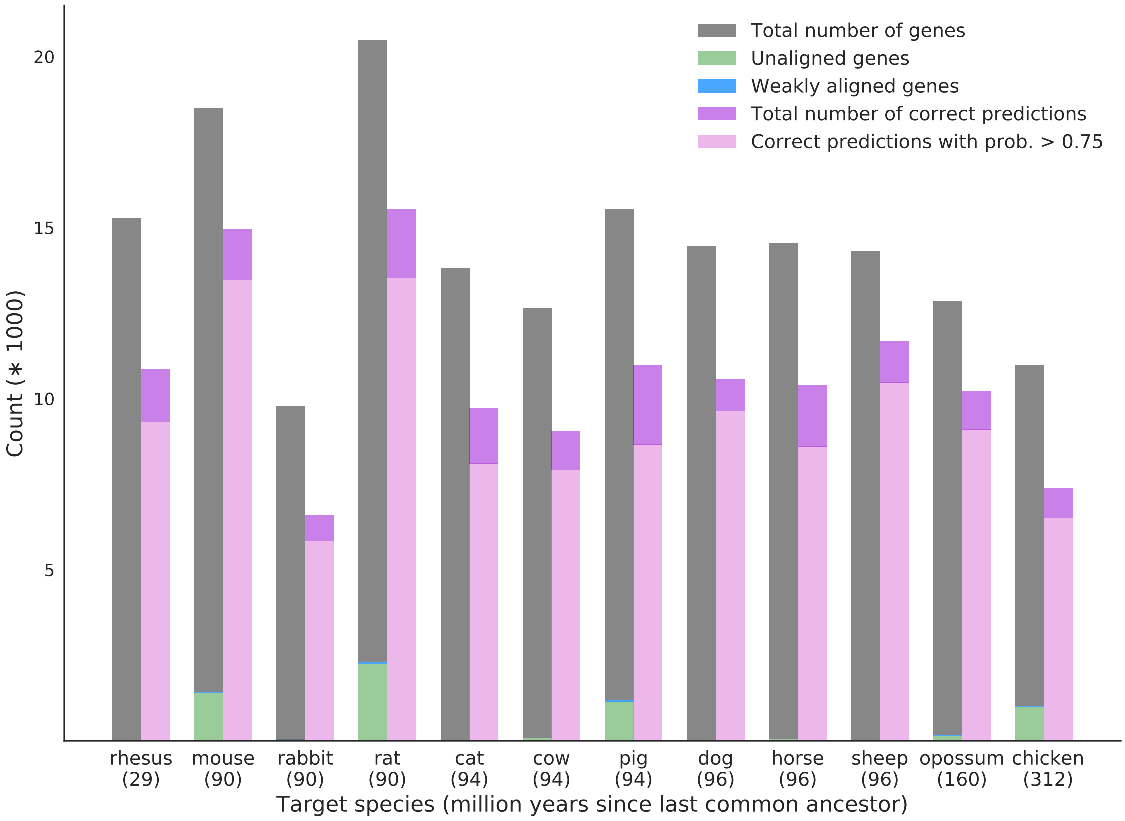


Figure 5.13: Epigenome-based prediction of gene expression using the human reference: for each species, bar plots show the total number of genes in the genome, the number of unaligned or weakly aligned genes relative to the reference species, and the average number of genes being correctly predicted by the epigenome-based approach in total, and after setting a threshold on the predicted class probability of > 0.75. Target species are sorted by evolutionary distance to the human reference along the x-axis (million years since last common ancestor indicated in parentheses).

of repetitive DNA elements (genomic duplications, satellite repeats, long terminal repeats, LINE repeats) as well as regions characterized by the heterochromatin mark H3K9me3 (Appendix C.1, Table S3). These results indicate that cross-species epigenome transfer is not well suited for analyzing species-specific gene families and repetitive heterochromatin regions.

Finally, we investigated whether the evolutionary age of individual genes may be associated with the accuracy of the epigenome-based predictions. Using a recently published dataset of gene age annotations [142], we found that genes whose expression status was predicted incorrectly tend to have a younger evolutionary age than those for which the expression status was predicted correctly (Figures 5.14 and C.8). For example, the group of genes specific to mammals contains 52% more incorrectly predicted genes than would be expected based on the background distribution of gene ages; in contrast, the much larger group of genes shared across eukaryotes contains 7% more correctly predicted genes compared to expectation.

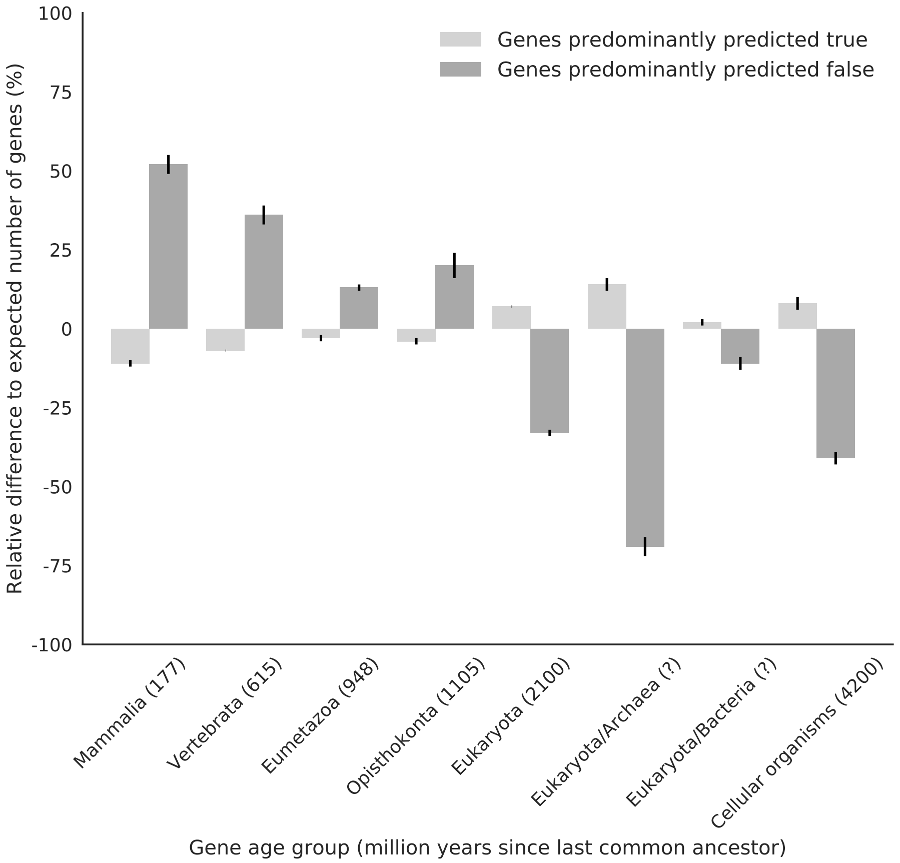


Figure 5.14: Association between gene age and epigenome-based prediction performance: genes were stratified by their tendency to be predicted correctly or incorrectly by the epigenome–based approach and labeled with their annotated age. Predictions were made in human using mouse as reference species. Bar heights indicate relative difference to the expected number of genes in each age group (percent values are shown for comparability across species). One standard deviation estimated by 1,000 bootstrap iterations is indicated as error bars. Numbers in parentheses indicate divergence time in million years relative to the family of Hominidae (great apes including human). Because of horizontal gene transfer, no divergence time estimates are given for the archaea and the bacteria group. The number of expected genes in each group is derived from the prior distribution for the gene age label.

5.4 Discussion

Epigenome profiling in the wet lab is costly and labor-intensive, and comprehensive tissue-specific epigenome resources are currently available for only a few species. Here we explored cross-species extrapolation of epigenome data as a new approach that utilizes the large catalogs of human and mouse reference epigenomes to support research in non-model organisms.

We have implemented a basic method for transferring epigenomes across species based on whole-genome alignments, which we applied and evaluated in two complementary ways. First, by comparing transferred to measured epigenome data for matched cell types between human and mouse, we collected evidence indicating that the cell-type specific nature of the epigenome is still detectable after cross-species transfer. Moreover, by correlating our transferred epigenome data to the external validation data by Villar et al., we gathered more evidence suggesting that a considerable portion of the epigenetic signal is successfully transferable between species, and could be put to use in subsequent computational analyses. Second, as a proof of concept example for such an analysis, we combined our cross-species epigenome transfer with epigenome-based prediction of gene expression. We could validate our gene expression predictions in a range of target species and observed generally high prediction accuracies. Since the cross-species epigenome transfer retained the cell-type specific regulatory pattern in the data, we could show that the predicted gene expression profiles were likewise tissue-specific. This is an important aspect in the cross-species analysis of gene regulation.

We observed broadly consistent results across twelve mammalian and one avian species included in our study, which span a spectrum of 20 million (mouse–rat) to 312 million (human–chicken and mouse–chicken) years since the last common ancestor. While it may be surprising that there is not a more pronounced decrease in prediction performance as evolutionary distances increase, all of the species included in our study share highly conserved epigenetic machinery, and, apparently, there is sufficient conservation in gene promoters and gene bodies to exploit this fact. Moreover, we found that evolutionarily old genes were predicted with higher accuracy than more recently evolved genes, which seems to be in line with the literature [189]. We assume that this factor is likely to contribute to the robustness of predictions across a wide evolutionary range.

Our method does not have major limitations that would hinder its broad application, except that it requires a reference genome for all included species. The more complete and more accurate the reference genome assemblies of the target species, the higher will be the quality of whole genome alignments, and, presumably, the performance of our method. Nevertheless, because our method transfers epigenome data between locally aligned regions, it is not restricted to high-quality genomes and might also cope reasonably well with initial, more fragmented genome assemblies. If this is successfully tested, cross-species epigenome transfer could indeed hint at new ways of investigating previously understudied non-model organisms.

CHAPTER 6

Conclusion and Perspectives

Modern biological research in the field of epigenomics is advancing at a fast pace, driven by technological progress in areas such as single-cell sequencing [45, 83, 222], and by setting ambitious goals such as creating the first complete atlas of all cell types in the human body [202]. The way to far-reaching goals is nevertheless studded with numerous individual studies, where each new generation of scientists builds upon the accomplishments of the previous one — or dismantles them in light of new evidence.
The scrutiny required to corroborate new scientific hypothesis is particularly laborious in emerging fields such as epigenomics. The work presented in Chapter 3 focused on reducing this workload for bioinformaticians, and on developing dissemination strategies for documented computational pipelines. To that end, a metadata model has been developed and published in *Oxford Database* [64] that is tailored to the needs of collaborative research consortia. This metadata model aims at limiting the amount of required manual work and at automating other tasks such as detecting errors and validating metadata records. The entire setup was successfully tested as part of the daily work routines in the DEEP consortium and, due to secondary features integrated over time, could actually provide more than the initially planned services. The resource created at doi.org/10.17617/1.2W provides the documentation of the DEEP pipelines in an online accessible, shareable and citable manner and is referenced on the official IHEC data portal [36].

After discussing standardized and presumably reproducible procedures for epigenome data processing (Chapter 3), an important next step is to enable scientists to compare biological samples in a meaningful way, e.g., to investigate healthy or disease cellular phenotypes. To that end, Chapter 4 introduced SCIDDO [65], a versatile and fast tool for the differential analysis of chromatin state segmentation maps. SCIDDO uses quantitative scores to measure chromatin state (dis-) similarity, an approach new to differential chromatin analysis. This new approach provides a potential advantage over competing methods, as it permits to explore varying notions of chromatin state similarity depending on the research question at hand. Moreover, SCIDDO does not require a large number of samples for the statistical evaluation, making it a suitable tool for the comparative analysis of small sample groups. In an exemplary study on a selected set of DEEP samples, SCIDDO's usefulness in identifying differential chromatin domains between even closely related cell types could be demonstrated. The identified differential chromatin domains are located in functionally plausible regions, e.g., in differentially expressed genes or in various regulatory elements, and we thus concluded that SCIDDO is a valid addition to existing software solutions for the analysis of chromatin data.

The principle of expanding biological knowledge by performing comparative analyses is pushed beyond intraspecies comparisons in Chapter 5. This chapter presented an exploratory study that crosses the species boundary between well-characterized model and understudied non-model species purely *in silico*. Our results show that cross-species transfer of epigenome data is possible among mammalian (and avian) species, and that the transferred epigenomes not only retain tissue specificity but also enable tissue-specific prediction of gene expression. We thus concluded that the tissue-specific links between epigenome profiles and gene expression are well conserved across the analyzed species. Although cross-species transfer of epigenome data is not meant to replace experimental data, our initial results suggest that *in silico* approaches could complement experimental analysis by providing access to a larger number of (non-model) species and tissues at essentially no additional cost. Hence, bioinformatic approaches for cross-species epigenome analysis suggest new use cases for chromatin data in comparative studies.

Perspectives

Issues arising from a lack of standardized data handling and processing procedures are still prevalent in the field of epigenomics. These issues manifest themselves, e.g., in the course of ongoing collaborative efforts such as the IHEC "EpiMAP"[1] project. The "EpiMAP" project aims at an integrative analysis of all reference epigenomes produced by the initiatives organized under the IHEC umbrella (DEEP, BLUEPRINT etc.), and faces many challenges related to data wrangling and metadata processing[2]. Although it is undoubtedly too early, and presumably generally unrealistic, to expect a seamless integration of data from a considerable number of different sources, projects such as "EpiMAP" highlight the importance of consistent metadata models, and of replicable analysis setups to reduce the workload during later data integration tasks. IHEC members continue their efforts to increase standardization, e.g., concerning the definition of quality control metrics for wet lab experiments[3], or the use of uniform data processing pipelines[4] at all collaborating partner institutions. The latter point implies that some of the work for the design and setup of computational analysis pipelines in DEEP has now been superseded, but this is a positive development indicating that the field is moving toward common grounds. Such gradual improvements can be considered the necessary steps to establish a solid basis for reproducible research in epigenomics.

For SCIDDO, a direct comparison with a competing method could not be realized due to the methodological differences and the dataset we used for our study (see Chapter 4, Sections 4.1

[1] Working title defined and used in the IHEC Integrative Analysis work group.

[2] Personal communication within the context of the regular IHEC Integrative Analysis work group conference calls.

[3] github.com/IHEC/ihec-assay-standards: IHEC Assay Standards work group code repository hosting reference scripts that implement computational steps to derive quality control metrics for experiments such as ChIP-seq.

[4] For example, at the time of writing, the ENCODE ChIP-seq pipeline v2 is considered the IHEC reference: github.com/ENCODE-DCC/chip-seq-pipeline2

and 4.3.9). Hence, future work may evaluate the differences arising from using a binary “match/mismatch” chromatin state scoring as opposed to the default quantitative scoring available in SCIDDO. However, drawing general conclusions from such a comparison may not be trivial because the lack of a genome-wide gold standard defining true regions of differential chromatin presents a challenge to deciding on the appropriateness of the respective chromatin state scoring in all genomic contexts. Similarly, reproducing the reported results using a different chromatin state segmentation tool to generate the input data, e.g., EpiCSeg [154], could be easily realized but differences in the output may be hard to interpret since different software builds on different ChIP-seq modeling assumptions (Chapter 2, Section 2.4.2). Consequently, it would be surprising to see virtually identical results, but general trends such as a formation of differential chromatin domains in gene bodies of differentially expressed genes should hold nevertheless.

Apart from consolidating work, exploring different notions of chromatin state (dis-) similarity by developing more, potentially data-derived, scoring schemes is an intriguing possibility. Specifically, it is an open question if, for a given state segmentation model, generally applicable scoring schemes could be created that are sensitive to the degree of cellular relatedness (in some sense analogous to scoring matrices for biological sequence analysis [182]). In the affirmative case, this could enable a more precise analysis of closely related cell types, which could be relevant for examining chromatin changes during subsequent steps of cellular differentiation. Similar considerations may be relevant when different types of (regulatory) genomic regions are in the focus of the differential chromatin analysis. By and large, the results presented in Chapter 4 have a gene-centric view, probably covering a substantial number of regions with pronounced differences on the chromatin level between cell types. However, fine-grained enhancer dynamics are still poorly understood on the global scale [35, 37, 185], and it is certainly worthwhile to test if a suitable combination of chromatin state segmentation model and scoring scheme can be devised to investigate (smaller-scale) chromatin changes in enhancer regions.

In the narrow context of this thesis, the appropriate handling of the so-called background chromatin state is of lesser relevance, because none of the work presented here could explicitly help dealing with technical artifacts that may cause the lack of a signal in certain genomic regions. From a more general perspective, a presumably context-dependent scoring of the background state is likely required to comprehensively characterize chromatin dynamics between cellular phenotypes within a species.

Knowledge about common patterns of chromatin state changes within one species could prove useful for improving cross-species studies that rely on epigenome transfer as in Chapter 5. Such additional information may help to broaden the applicability of epigenome transfer beyond coding genes; in particular, this will be a challenge for regulatory elements such as enhancers that show a limited sequence conservation. Nevertheless, it seems feasible that cross-species epigenome transfer and prediction can help address the current scarcity of epigenome data for non-model organisms. A more in-depth comparison of predicted and experimentally measured epigenome data, ideally across a large panel of histone marks, cell types and species, would be interesting for locating genomic regions in which the measurement deviates from the prediction. Such regions would be strong candidates for species-specific epigenome regulation and promising targets for in-depth biological investigation. Species-specific properties of the epigenome are not only relevant for the endeavor of understanding epigenome evolution [149], but can reveal intriguing, and potentially medically relevant insights; a recent example is the discovered resistance to cellular reprogramming of the naked mole rat's epigenome, which may contribute to this species' exceptional resistance to cancer [236].

On the more computational side, it would be interesting to test if the epigenetic information from several reference species could be combined during epigenome transfer or machine learning to increase the robustness of the approach. However, this would probably require larger and more biologically consistent datasets; a non-trivial task given that characteristics such as (sample) age or feeding status are difficult to match across species. Moreover, it would be desirable to perform more in-depth work on a dataset restricted to purified cell types, i.e., to exclude whole-tissue samples that likely contain a mix of different cells and thus a mix of biological signals.

Additionally, one could consider explicitly modeling confounding factors. For example, the estimated gene age seems to be a potential candidate to perform a preliminary analysis along those lines. However, such extensions bear the risk of limiting the applicability of the computational approach if they introduce too strong dependencies on genomic annotations that have to be available for all species considered.

At this early stage, it seems unlikely that pure bioinformatic approaches to cross-species epigenomics will be sufficient to derive general principles of epigenome evolution — if there are any at all. On the other hand, a deeper understanding of epigenome dynamics within a species may allow for dropping the proxy of "sequence conservation suggests epigenome conservation" at some point, which could then motivate entirely different bioinformatic approaches to *in silico* cross-species epigenomics.

CHAPTER A

Appendix: Reproducibility in Computational Research

A.1 Additional Material

A.1.1 Example of a Galaxy workflow specification

The following example of a published Galaxy workflow document[1] represents a ChIP-seq analysis pipeline that implements the main steps of read mapping and peak calling, and is roughly comparable to the respective DEEP pipelines. The JSON document is structured similar to DEEP process documents beginning with a header containing information such as the workflow name (Listing A.1), followed by the specification of the input dataset (Listing A.2) and the individual analysis steps of the pipeline. The example step presented in Listing A.3 is the peak calling using the MACS [268] software. Noteworthy differences between DEEP process documents and Galaxy workflows are, e.g., the inclusion of layout information for proper rendering in the Galaxy software (see keyword “position” in lines 18 and 59 in Listings A.2 and A.3); the explicit specification of software repositories (tool sheds in Galaxy lingo, see lines 64ff in Listing A.3) and the specification of input data as extra pipeline step instead of a separate section (see Listing A.2).

Listing A.1: Lines 1 to 5 of a Galaxy ChIP–seq workflow: document header

```
1  {
2     ”a_galaxy_workflow”: ”true”,
3     ”annotation”: ””,
4     ”format−version”: ”0.1”,
5     ”name”: ”ChIP(Helin)−2nd step”,
```

[1] Source: usegalaxy.org/workflows/list_published, downloaded on 2018-05-13. The workflow was last updated on 2015-04-28 and has the universally unique identifier (UUID) “447c66c4-a965-4a17-a230-95fef65788de”. For layout reasons, white space and indentation have been adapted to fit the page width.

Listing A.2: Lines 6 to 33 of a Galaxy ChIP–seq workflow: input dataset

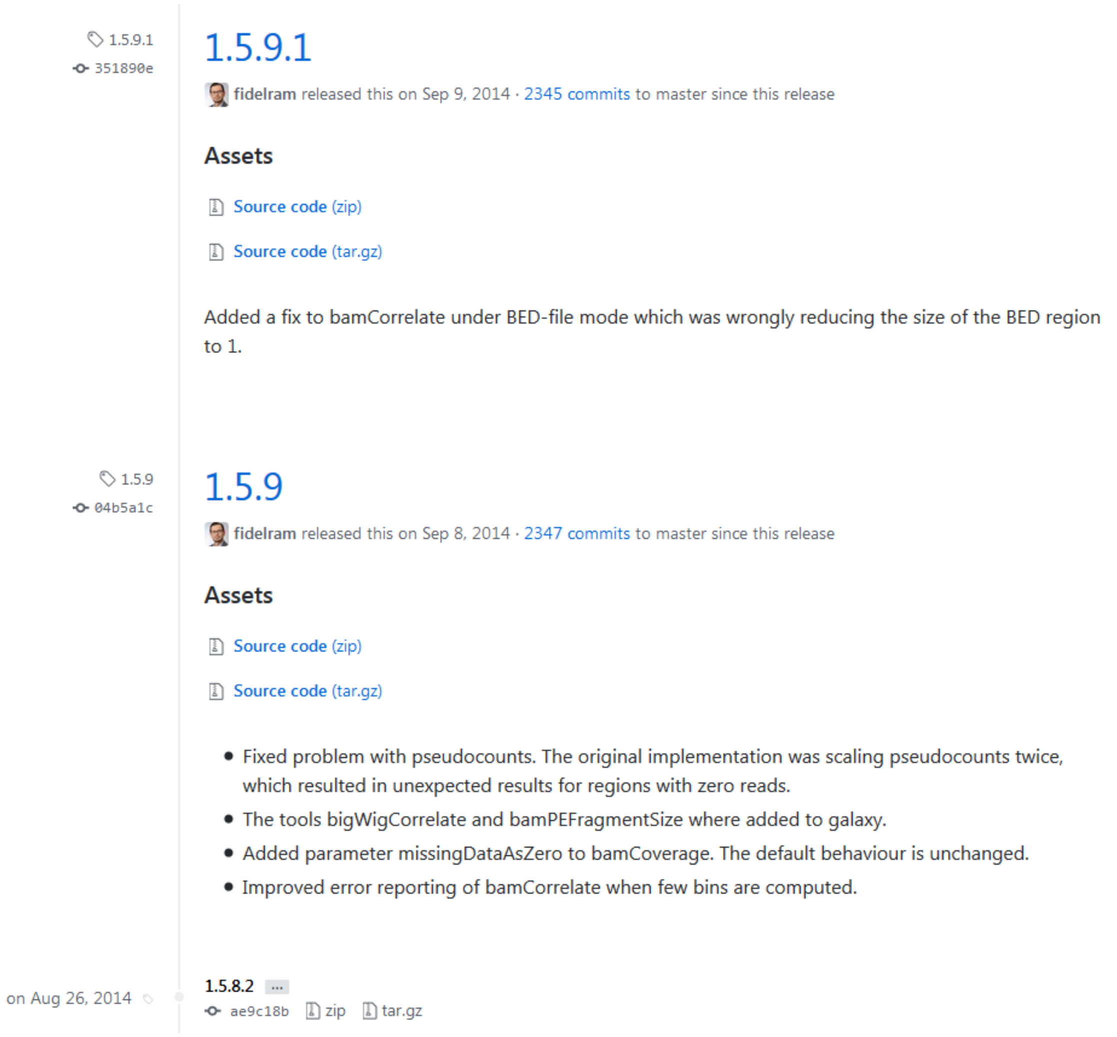
```
    ”steps”: {
        ”0”: {
            ”annotation”: ””,
            ”content_id”: null,
            ”errors”: null,
            ”id”: 0,
            ”input_connections”: {},
            ”inputs”: [
                {
                    ”description”: ””,
                    ”name”: ”ChIP−seq against
                             protein of interest”
                }
            ],
            ”label”: null,
            ”name”: ”Input dataset”,
            ”outputs”: [],
            ”position”: {
                ”left”: 211,
                ”top”: 200
            },
            ”tool_id”: null,
            ”tool_state”: ”{\”name\”: \”ChIP−seq against
                                    protein of interest\”}”,
            ”tool_version”: null,
            ”type”: ”data_input”,
            ”uuid”: ”None”,
            ”workflow_outputs”: []
        },
```

Listing A.3: Lines 234 to 345 of a Galaxy ChIP–seq workflow: peak calling

```
        ”6”: {
            ”annotation”: ””,
            ”content_id”: ”toolshed.g2.bx.psu.edu/repos/
                           devteam/macs/
                           peakcalling_macs/1.0.1”,
            ”errors”: null,
            ”id”: 6,
            ”input_connections”: {
                ”input_type|input_chipseq_file1”: {
                    ”id”: 4,
                    ”output_name”: ”output1”
                },
                ”input_type|input_control_file1”: {
```

```
                ”id”: 1,
                ”output_name”: ”output”
            }
        },
        ”inputs”: [
            {
                ”description”: ”runtime parameter
                                for tool MACS”,
                ”name”: ”gsize”
            },
            {
                ”description”: ”runtime parameter
                                for tool MACS”,
                ”name”: ”pvalue”
            }
        ],
        ”label”: null,
        ”name”: ”MACS”,
        ”outputs”: [
          {
            ”name”: ”output_bed_file”,
            ”type”: ”bed”
          },
          {
            ”name”: ”output_xls_to_interval_peaks_file”,
            ”type”: ”interval”
          },
          {
            ”name”: ”output_xls_to_interval_negative
                     _peaks_file”,
            ”type”: ”interval”
          },
          {
            ”name”: ”output_treatment_wig_file”,
            ”type”: ”wig”
          },
          {
            ”name”: ”output_control_wig_file”,
            ”type”: ”wig”
          },
          {
            ”name”: ”output_extra_files”,
                ”type”: ”html”
            }
        ],
        ”position”: {
                ”left”: 1228,
                ”top”: 452
        },
        ”post_job_actions”: {},
        ”tool_id”: ”toolshed.g2.bx.psu.edu/repos/devteam/
                                macs/peakcalling_macs/1.0.1”,
        ”tool_shed_repository”: {
                ”changeset_revision”: ”ae2ec275332a”,
                ”name”: ”macs”,
                ”owner”: ”devteam”,
                ”tool_shed”: ”toolshed.g2.bx.psu.edu”
        },
        ”tool_state”:
          ”{\”gsize\”: \”{\\\”__class__\\\”:
                  \\\”RuntimeValue\\\”}\”,
              \”tsize\”: \”\\\”42\\\”\”,
              \”nolambda\”: \”\\\”false\\\”\”,
              \”__page__\”: null,
              \”input_type\”:
              \”{\\\”input_control_file1\\\”: null,
              \\\”input_chipseq_file1\\\”: null,
              \\\”__current_case__\\\”: 1,
              \\\”input_type_selector\\\”:
                  \\\”single_end\\\”}\”,
              \”__rerun_remap_job_id__\”: null,
              \”diag_type\”:
                  \”{\\\”__current_case__\\\”: 1,
              \\\”diag_type_selector\\\”:
                  \\\”no_diag\\\”}\”,
              \”wig_type\”:
              \”{\\\”wig_type_selector\\\”:
              \\\”wig\\\”,
              \\\”wigextend\\\”: \\\”200\\\”,
              \\\”__current_case__\\\”: 0,
              \\\”space\\\”: \\\”10\\\”}\”,
              \”xls_to_interval\”: \”\\\”true\\\”\”,
              \”experiment_name\”:
                  \”\\\”MACS in Galaxy\\\”\”,
                \”bw\”: \”\\\”300\\\”\”,
                \”futurefdr\”: \”\\\”false\\\”\”,
                \”nomodel_type\”:
                \”{\\\”nomodel_type_selector\\\”:
                \\\”create_model\\\”,
                \\\”__current_case__\\\”: 1}\”,
                \”mfold\”: \”\\\”32\\\”\”,
                \”lambdaset\”: \”\\\”1000,5000,10000\\\”\”,
                \”pvalue\”: \”{\\\”__class__\\\”:
                  \\\”RuntimeValue\\\”}\”}”,
            ”tool_version”: ”1.0.1”,
            ”type”: ”tool”,
            ”uuid”: ”None”,
            ”workflow_outputs”: []
        },
```

A.2 Additional Figures



1.5.9.1
351890e

1.5.9.1

fidelram released this on Sep 9, 2014 · 2345 commits to master since this release

Assets

Source code (zip)

Source code (tar.gz)

Added a fix to bamCorrelate under BED-file mode which was wrongly reducing the size of the BED region to 1.

1.5.9
04b5a1c

1.5.9

fidelram released this on Sep 8, 2014 · 2347 commits to master since this release

Assets

Source code (zip)

Source code (tar.gz)

- Fixed problem with pseudocounts. The original implementation was scaling pseudocounts twice, which resulted in unexpected results for regions with zero reads.
- The tools bigWigCorrelate and bamPEFragmentSize where added to galaxy.
- Added parameter missingDataAsZero to bamCoverage. The default behaviour is unchanged.
- Improved error reporting of bamCorrelate when few bins are computed.

on Aug 26, 2014

1.5.8.2 ...
ae9c18b zip tar.gz

Figure A.1: Github changelog of the deepTools software suite: screenshot of the Github changelog entries for the deepTools releases 1.5.8.2, 1.5.9 and 1.5.9.1 that followed on release 1.5.7. Source: github.com/deeptools/deepTools

CHAPTER B

Appendix: Fast Detection of Differential Chromatin Domains with SCIDDO

B.1 Additional Figures

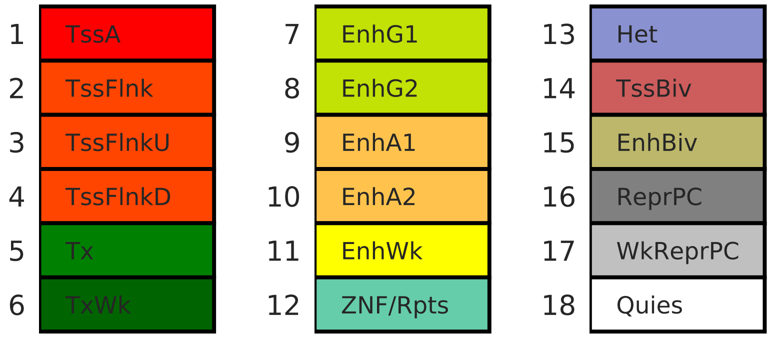


Figure B.1: CMM18 state mnemonics and colors: mnemonics and colors for the 18 chromatin states of the ChromHMM CMM18 model provided by the REMC. See Table B.1 for more detailed state descriptions.

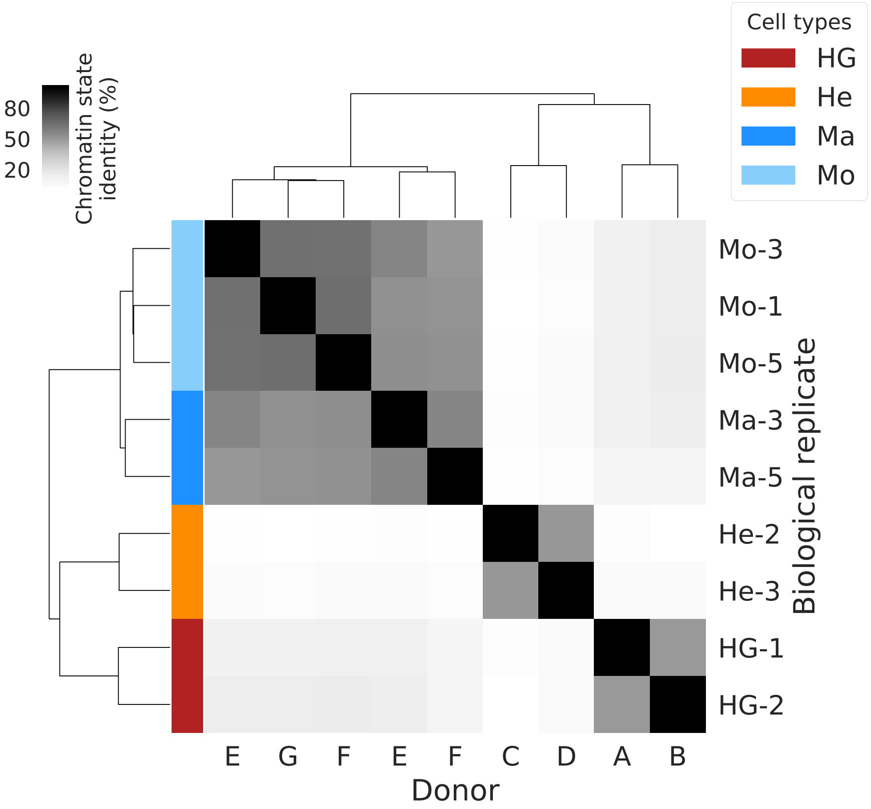


Figure B.2: Clustering of samples based on chromatin state identity: fraction of identically assigned chromatin state bins was computed for all sample pairs based on the segmentation produced by the CMM18 model. The resulting similarity matrix was used for a hierarchical clustering of the samples with average linkage and Euclidean distance. Donors were arbitrarily labeled A–G (x-axis) based on the DEEP sample metadata (the given labels are otherwise not informative).

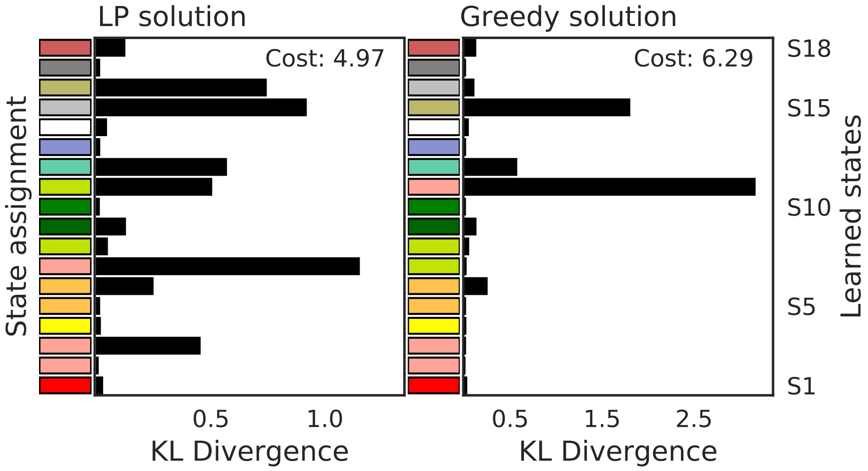


Figure B.3: Matching of newly learned and predefined chromatin states: the 18 chromatin states $S1, \ldots, S18$ of the NEW18 model are arranged in sorted order from bottom to top (rightmost y-axis) and the matched states of the CMM18 model are shown as colored blocks (see Figure B.1) for the LP (left) and the greedy (right) solution of the assignment problem defined in Section 4.2.3. The Kullback-Leibler divergence between the matched states is depicted on the x-axis. Note that the color of the CMM18 states 2–4 has been changed to “salmon” to make the TSS flanking states distinguishable from the active TSS state (bright red).

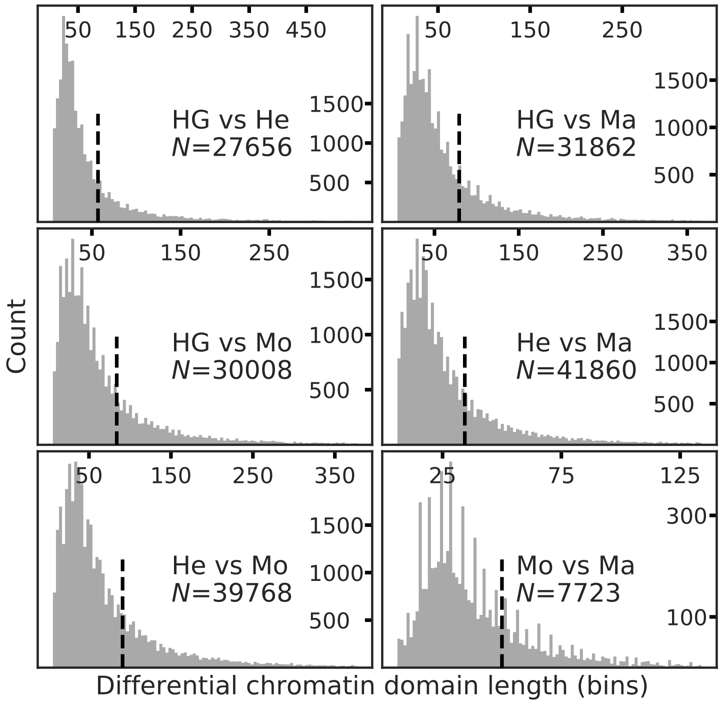
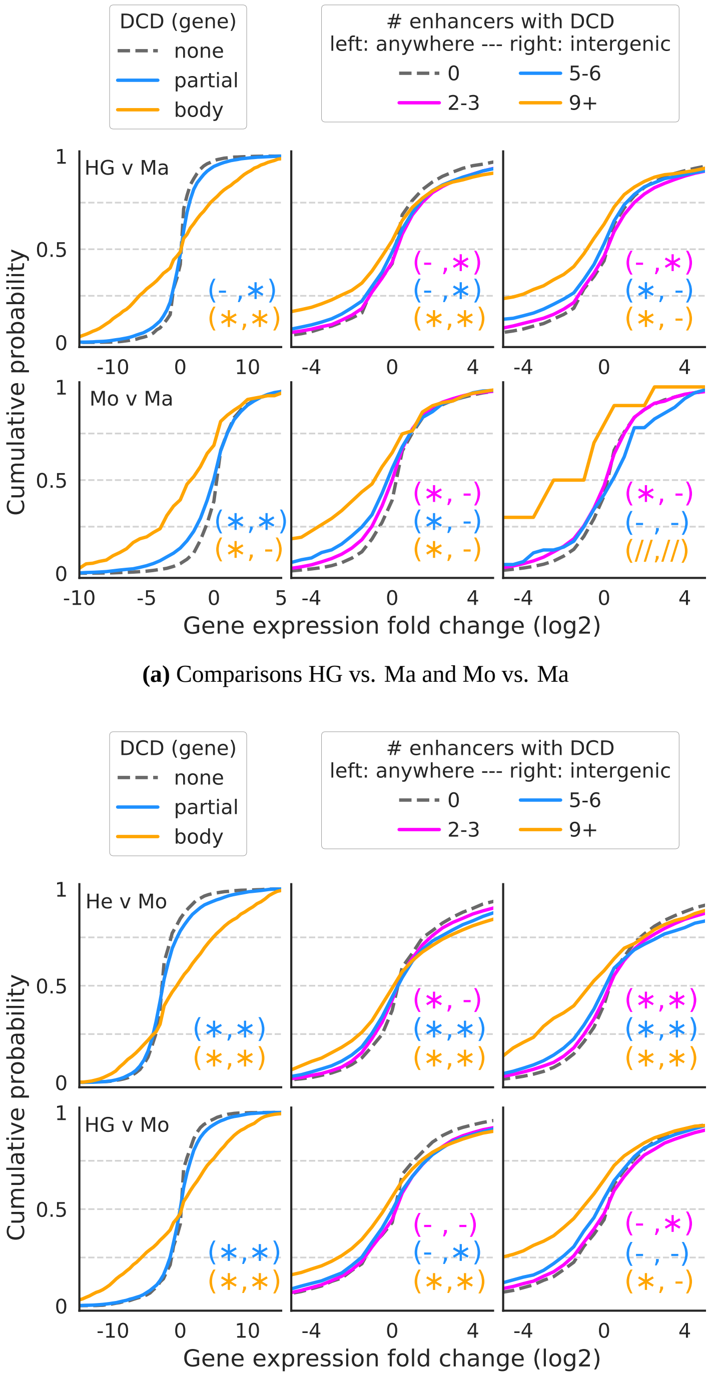


Figure B.4: DCD length distribution: length distribution for the lower 97.5% (truncated for visualization) of all identified differential chromatin domains for the six sample group comparisons. The vertical line (dashed) marks the 75th percentile of the data. The DCD length is given in genomic bins à 200 bp (x-axis). N: total number of identified DCDs in the respective comparison.



(a) Comparisons HG vs. Ma and Mo vs. Ma

(b) Comparisons HG vs. Mo and He vs. Mo

Figure B.5: DCDs overlapping gene bodies and enhancers affect gene expression: (left panels) genes were stratified by the amount of DCD overlap either covering more than 50% of the body (body; orange curve) or less than 50% of the body or the promoter region (partial; blue curve). Expression fold change of the genes in the respective groups is plotted along the x-axis within a restricted window for improved readability. Statistical significance of the difference in mean fold change of the groups relative to the no overlap group ("none", gray dashed line) was computed separately for negative and positive fold change genes using a two-sided Mann-Whitney-U test ("*" significant at $p < 0.01$, "-" not significant otherwise; "//" not enough data to compute statistic). (middle and right panels) the same analysis as for the gene body was performed, but here counting the number of intra- and intergenic enhancers (anywhere, middle) or only intergenic enhancers (right) per gene that overlap a DCD. Expression fold changes plotted within a restricted window for improved readability. Statistical significance assessed as before.

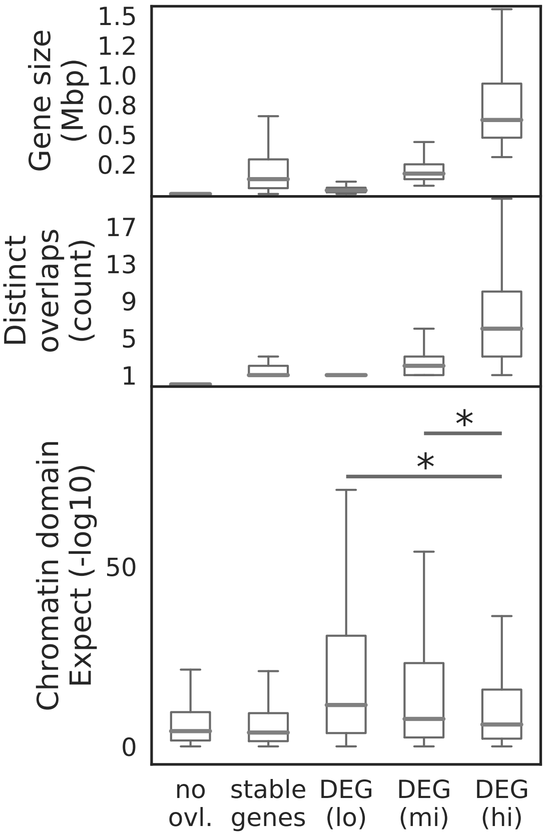


Figure B.6: E-value distribution in DEGs by gene body length: genes were stratified into four groups based on expression behavior (stable or differential) and their gene body length (shortest 40%, middle 40% and longest 20% of DEGs according to gene body length). Bottom: boxplots show distribution of E-values of all DCDs overlapping gene bodies in the respective groups aggregated over all sample comparisons. The no overlap group contains all E-values of DCDs not overlapping any gene. Middle: boxplots show distinct DCD overlaps per gene. Top: boxplots show gene body length distribution of all genes in the respective group. Differences in magnitude of E-values were assessed with a two-sided Mann-Whitney-U test and considered significant (*) at $p < 0.01$.

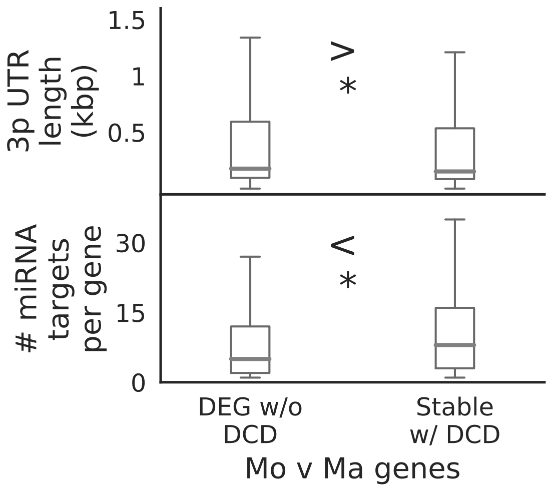


Figure B.7: Potential differences in post–transcriptional regulation of genes overlapping DCDs: DEGs w/o a DCD in their gene body were compared to the stably expressed genes with an overlapping DCD (N=760) in their gene body for the monocyte to macrophage comparison. Top: boxplots show distribution of 3p UTR length as annotated in Ensembl v78 for the genes in the respective groups. Bottom: boxplots show distribution of number of annotated miRNA targets per genes in the respective groups (TargetScan v7.2). Differences in magnitude between the two groups were assessed with a one-sided Mann-Whitney-U test (alternative less or greater as indicated) and considered significant (*) at $p < 0.01$.

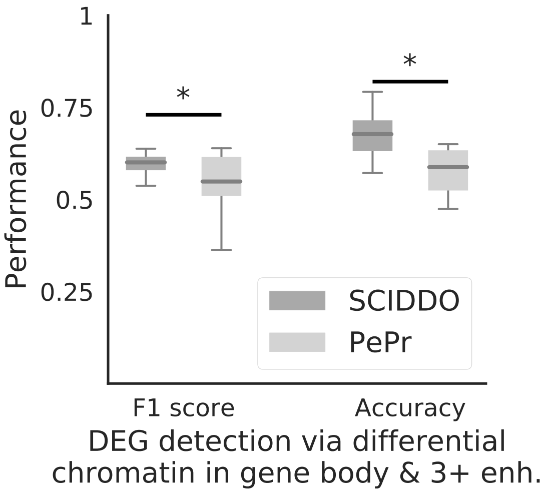


Figure B.8: SCIDDO shows more stable performance at detecting DEGs (G2): boxplots depict SCIDDO's and PePr's (light grey) performance of detecting DEGs quantified as F1 score (left) and as accuracy (right). Performance values are summarized over all sample group comparisons and for different thresholds on gene expression fold change (0.5, 1, 2 and 4) and on adjusted p-values (0.1, 0.05, 0.01 and 0.001) computed with DESeq2 to call DEGs. At least one DCD/differential H3K36me3 peak (PePr) in the gene body and at least three DCDs/differential H3K27ac peaks (PePr) in gene-associated enhancers were required for a DEG to be considered detected on the chromatin level. Differences in performance were assessed with a one-sided Mann-Whitney-U test and considered significant "*" at $p < 0.01$

B.2 Additional Tables

Table B.1: State numbers, mnemonics and concise descriptions of the chromatin states of the ChromHMM CMM18 model as provided by the REMC under egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html

Number	Mnemonic	Description
1	TssA	Active TSS
2	TssFlnk	Flanking TSS
3	TssFlnkU	Flanking TSS upstream
4	TssFlnkD	Flanking TSS downstream
5	Tx	Strong transcription
6	TxWk	Weak transcription
7	EnhG1	Genic enhancer1
8	EnhG2	Genic enhancer2
9	EnhA1	Active enhancer 1
10	EnhA2	Active enhancer 2
11	EnhWk	Weak enhancer
12	ZNF/Rpts	ZNF genes & repeats
13	Het	Heterochromatin
14	TssBiv	Bivalent/Poised TSS
15	EnhBiv	Bivalent enhancer
16	ReprPC	Repressed PolyComb
17	WkReprPC	Weak repressed PolyComb
18	Quies	Quiescent/Low

Table B.2: Runtime (in minutes of wall clock time) of individual SCIDDO commands executed in order from top to bottom to perform the differential analysis presented in Chapter 4. The runtime for the `scan` command refers to a single comparison of two versus two samples. The last `scan` command is provided as an example of the scaling behavior of SCIDDO (scanning for differential chromatin domains between the four liver and the five blood samples in the dataset). Note that the runtime includes I/O.

Command	CPU cores	Samples	Runtime (min)
convert	7	9	< 3
stats	7	9	< 1
score	1	n/a	< 1
scan	15	2 v 2	< 4
scan	15	4 v 5	< 7

Table B.3: Average Spearman correlation of E-values of all overlapping candidate regions identified in individual replicate comparisons. Rightmost column indicates the average percentage of unique candidate regions per comparison. Values in parentheses give ± 1 standard deviation for the respective statistic.

Group 1	Group 2	Spearman's ρ	Unique regions %
HG	He	0.67 (0.06)	10.85 (3.14)
HG	Ma	0.7 (0.04)	7.51 (3.6)
HG	Mo	0.73 (0.04)	3.92 (1.68)
He	Ma	0.68 (0.04)	5.99 (1.35)
He	Mo	0.7 (0.03)	3.27 (0.39)
Mo	Ma	0.7 (0.04)	17.68 (3.17)

Table B.4: Online supplementary files are hosted in the source repository of this thesis under “Supplement/diffchrom” at github.molgen.mpg.de/pebert/dissertation

Table	Filename	Timestamp
S1	supp_table_S1_dataset.tsv	2018–10
S2	supp_table_S2_expression.tsv	2018–10

CHAPTER C

Appendix: Epigenome-based Prediction of Gene Expression across Species

C.1 Additional Figures

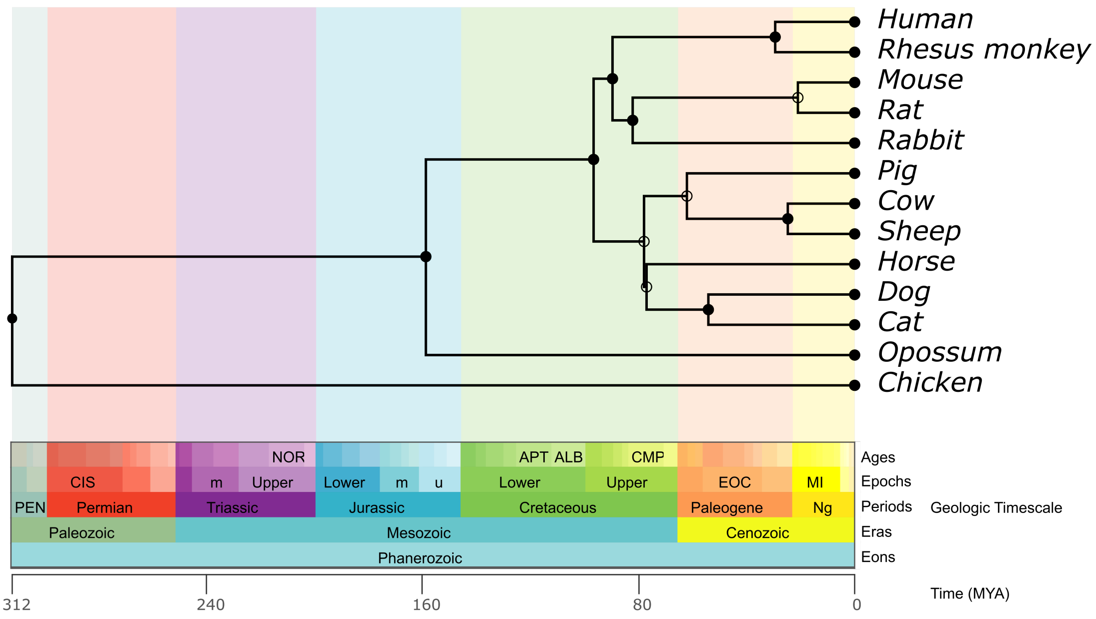


Figure C.1: Overview of species included in the cross-species study: phylogenetic tree showing evolutionary relationships among twelve mammalian and one avian species investigated in chapter 5. The tree was generated using the TimeTree web service [99]. Estimated time since last common ancestor between any two species is given in million years.

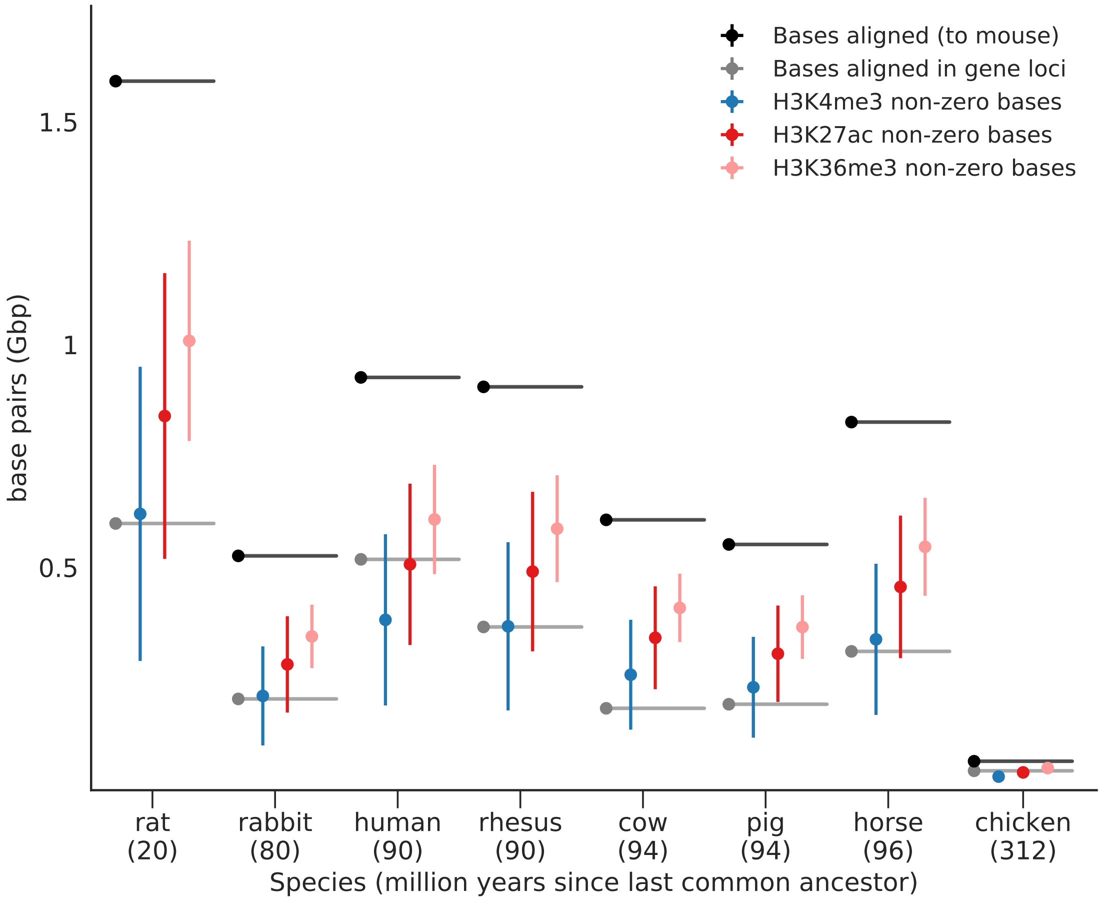
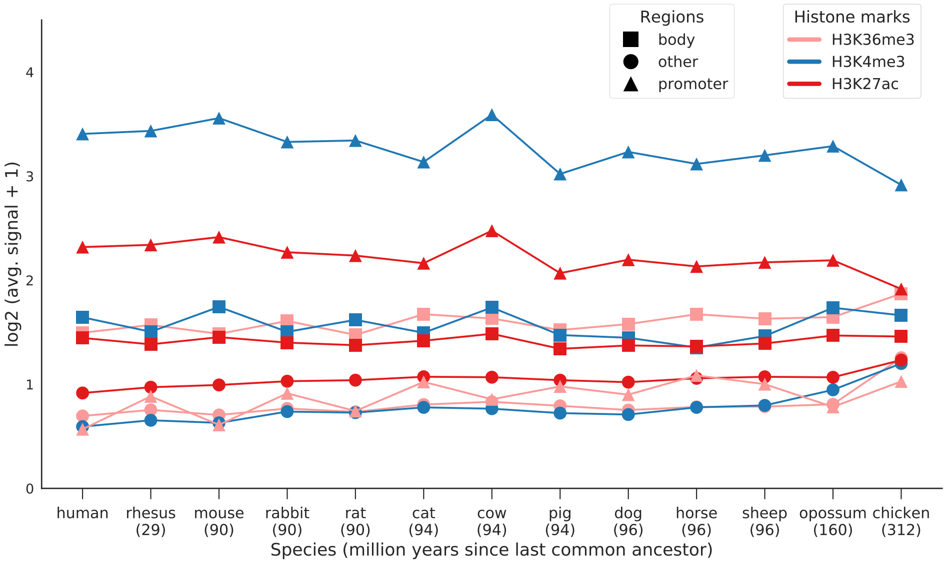
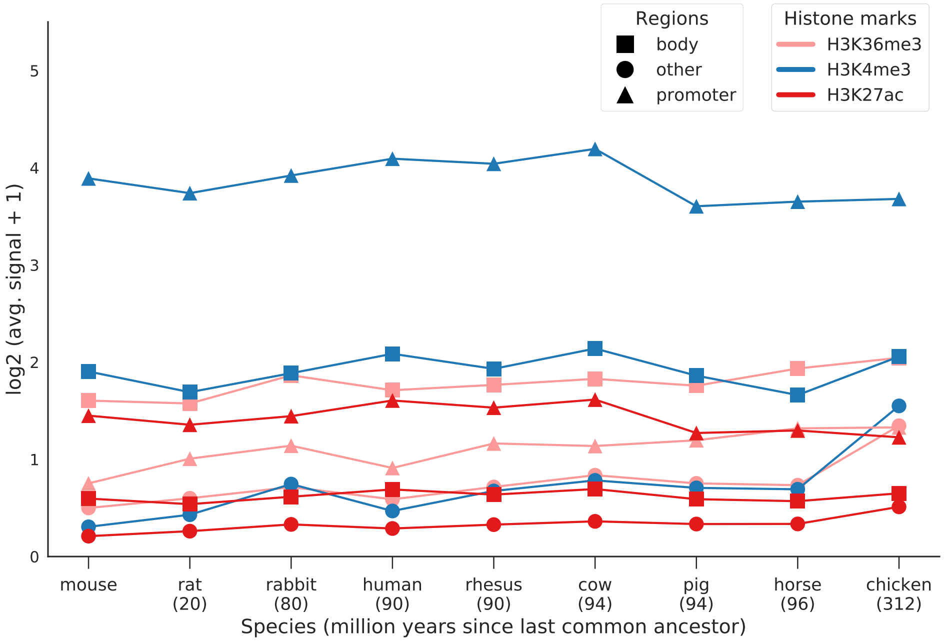


Figure C.2: Coverage of transferred epigenome profiles: colored points indicate the average number of base pairs with non-zero signal after cross-species transfer for each histone mark (error bars denote ± 1 standard deviation around the mean) between mouse as the reference species and twelve target species. The number of aligned bases genome-wide (black bars) and at gene loci (gene promoters and gene bodies combined, grey bars) are shown for comparison. Target species are sorted by increasing evolutionary distance to the mouse reference (x-axis, million years since last common ancestor indicated in parentheses).



(a) Human reference



(b) Mouse reference

Figure C.3: Strength of the transferred epigenetic signal: target species are sorted by increasing evolutionary distance (a) to the human reference and (b) to the mouse reference (x-axis, million years since last common ancestor). The average strength of the transferred histone signal in aligned regions in gene promoters, gene bodies and outside of genes ("other") is depicted as aggregation over all cell types (y-axis).

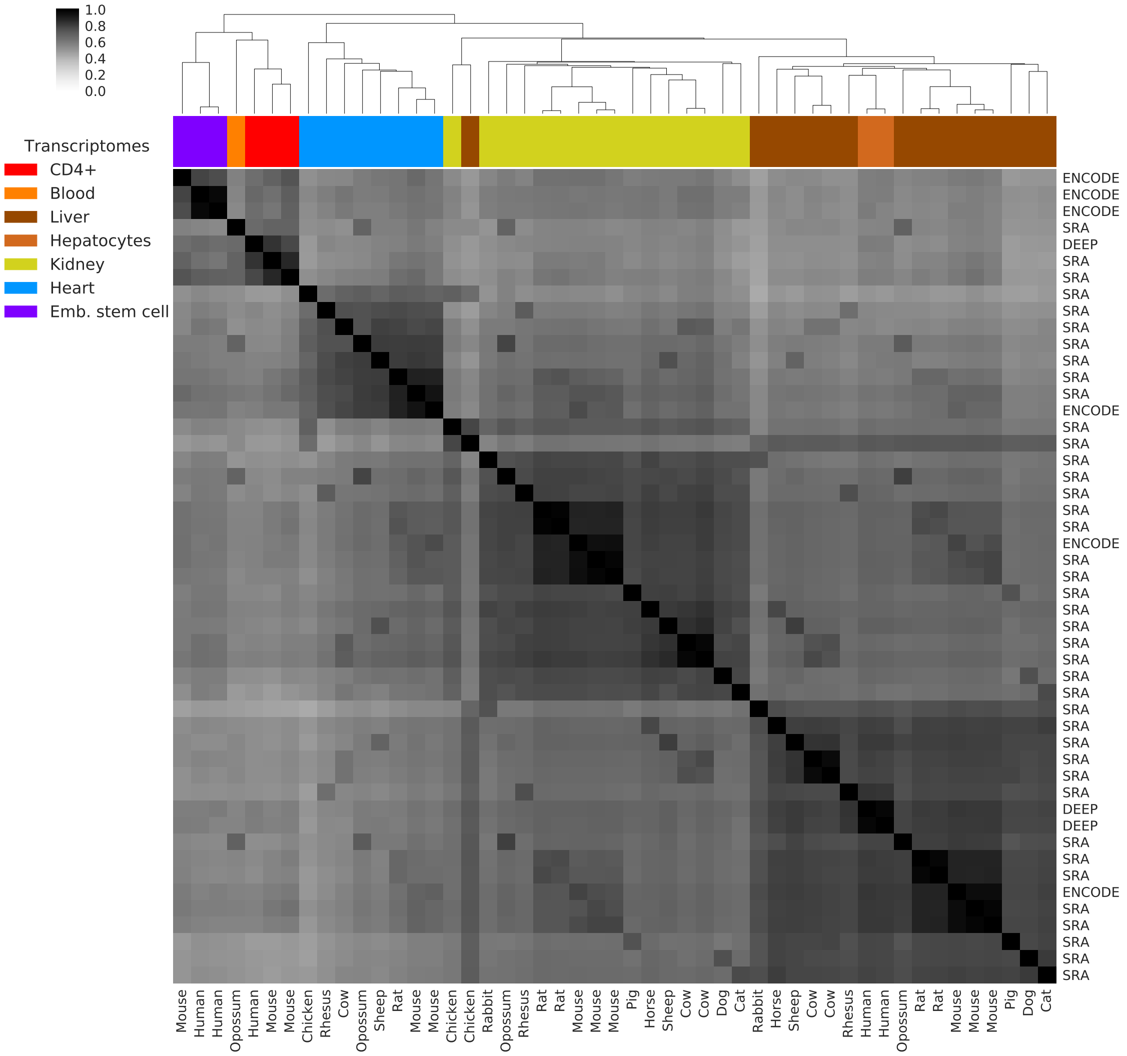
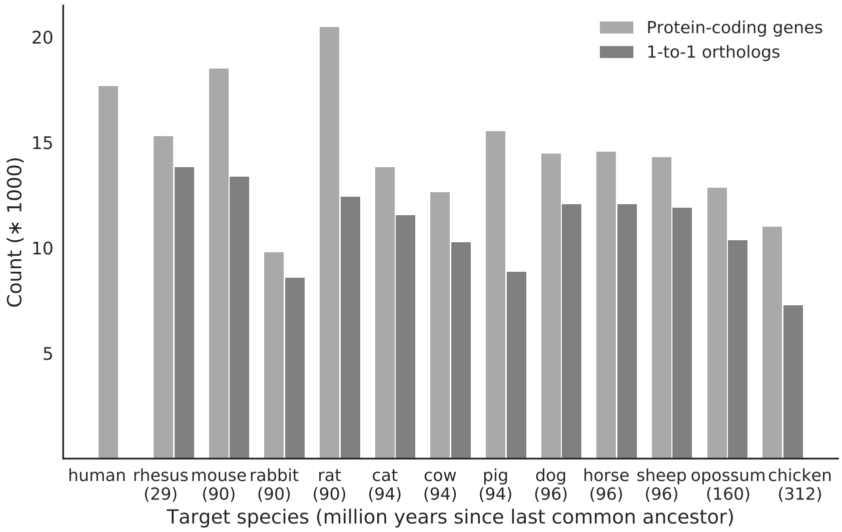
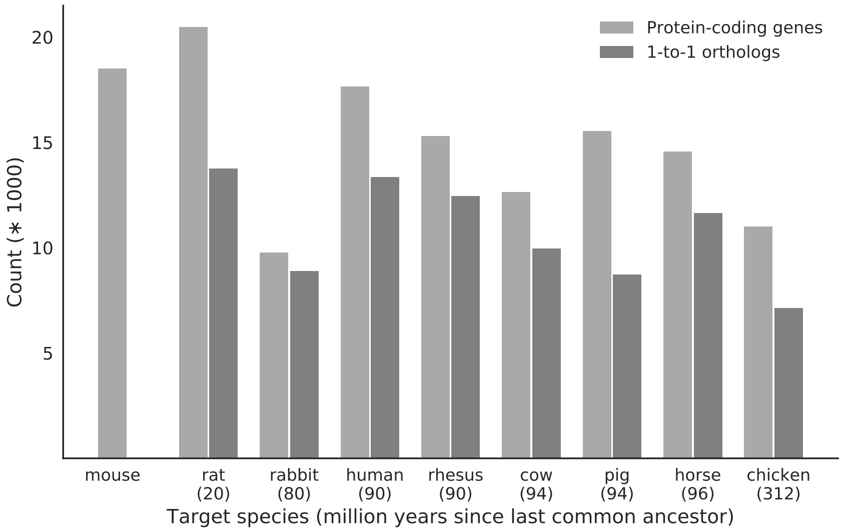


Figure C.4: Transcriptome data cluster by tissue of origin: hierarchical clustering of Pearson correlation coefficients between transcriptome profiles of 13 species used for gene expression prediction and model validation. Correlations were calculated across all 1-to-1 orthologous genes shared by all species in the dataset. Color bars at the top indicate tissue of origin. Columns are labeled with species of origin, and row labels indicate data source (project name or “SRA” for data downloaded from SRA/ENA; see Table C.1: Online Table S1).



(a) Human reference



(b) Mouse reference

Figure C.5: Number of annotated protein-coding and 1-to-1 orthologous genes: target species are sorted by increasing evolutionary distance (a) to the human and (b) to the mouse reference (x-axis, million years since last common ancestor). The number of annotated protein-coding genes located on autosomes and the number of 1-to-1 orthologs is depicted as height of the bars.

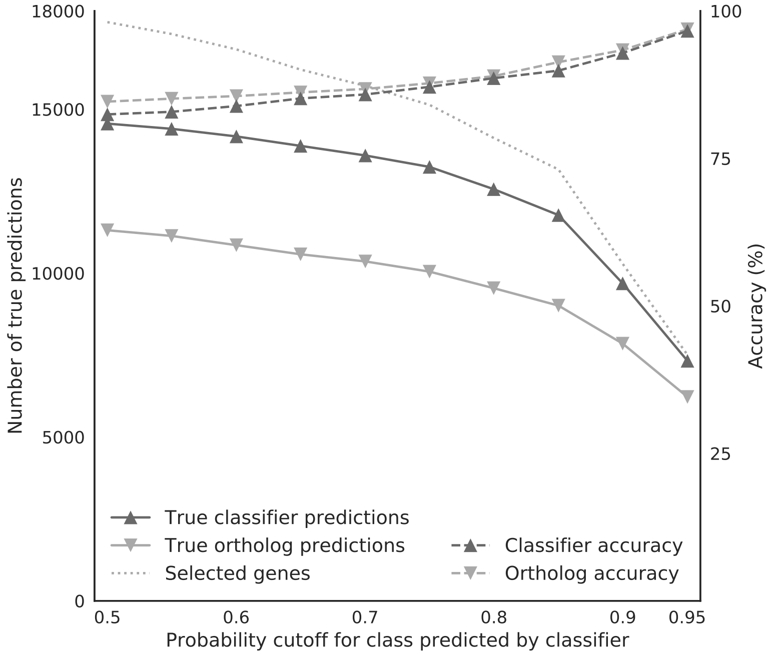


Figure C.6: Effect of thresholding on classifier class probability estimates: prediction performance of the epigenome-based approach (dark gray lines) and the orthology-based approach (light gray lines), as measured by the number of correct predictions (true positives and true negatives, left y-axis) and percent accuracy (right y-axis) for increasingly stringent cutoffs on the class probability estimates of the epigenome-based approach (x-axis, range from 0.5 to 0.95). The number of genes selected at each threshold is shown as dotted line. Results are shown for mouse as reference and human as target species.

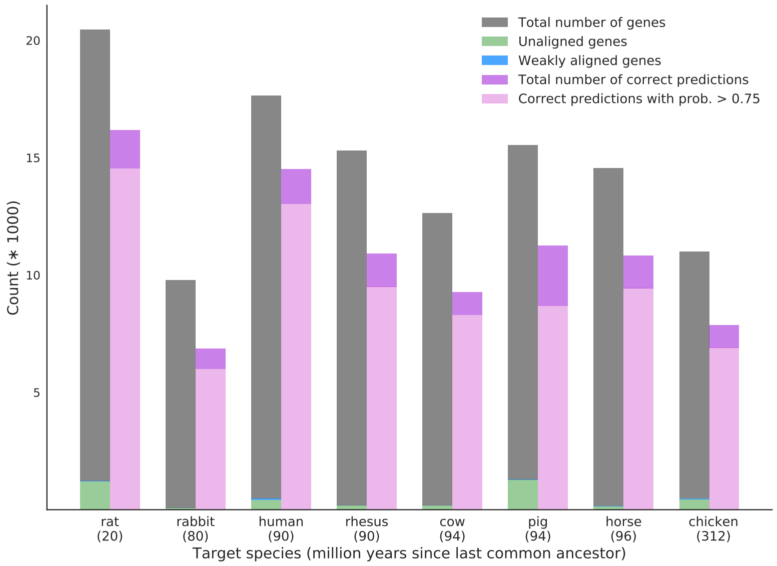
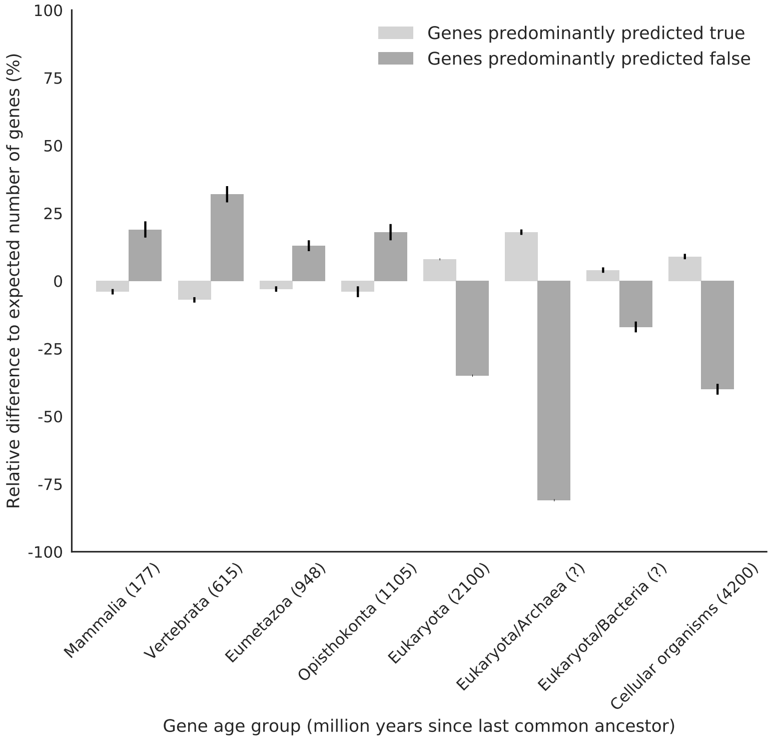
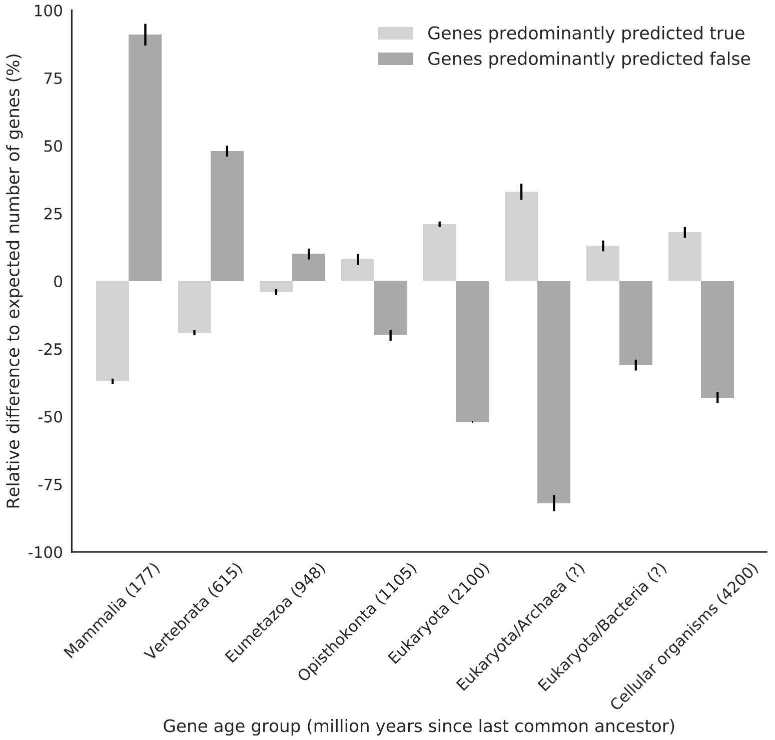


Figure C.7: Epigenome-based prediction of gene expression using the mouse reference: bar plots showing for each species the total number of genes in the genome, the number of unaligned or weakly aligned genes relative to the reference species, and the average number of genes correctly predicted by the epigenome-based approach in total and after setting a threshold on the predicted class probability of > 0.75. Target species are sorted by evolutionary distance to the mouse reference along the x-axis (million years since last common ancestor indicated in parentheses).



(a) Target species: mouse



(b) Target species: rhesus monkey

Figure C.8: Association between gene age and epigenome-based prediction performance: genes were stratified by their tendency to be predicted correctly or incorrectly by the epigenome-based approach and labeled with their annotated age. Predictions were made (a) in mouse using human as reference and (b) in rhesus monkey using human and mouse as reference species. Bar heights indicate relative difference to the expected number of genes in each age group (percent values are shown for comparability across species). One standard deviation estimated by 1,000 bootstrap iterations is indicated as error bars. Numbers in parentheses indicate divergence time in million years relative to the family of Hominidae (great apes including human). Because of horizontal gene transfer, no divergence time estimates are given for the archaea and the bacteria group. The expected number of genes in each group is derived from the prior distribution for the gene age label.

C.2 Additional Tables

Table C.1: Online supplementary files are hosted in the source repository of this thesis under “Supplement/crossspecies” at github.molgen.mpg.de/pebert/dissertation

Table	Filename	Timestamp
S1	Supp_Table_S1_data-sources.tsv	2018–07
S2	Supp_Tables_S2_LOLA_transfer.zip	2018–07
S3	Supp_Tables_S3_LOLA_wkalign.zip	2018–07
S4	Supp_Tables_S4_enrichR_wkalign.zip	2018–07

CHAPTER D

Appendix: List of (Co-) Authored Publications

Peer-reviewed publications and manuscripts under review

List in alphabetical order by first author name per year. List of authors is reduced to the first five authors for layout reasons. Items marked with “-1-” are first-author publications, “-2-” indicates co-first authorship, and “-n-” indicates contributing authorship.

2018:

-n- Fatemeh Behjati Ardakani, Kathrin Kattler, Karl Nordström, Nina Gasparoni, Gilles Gasparoni, et al. Integrative analysis of single-cell expression data reveals distinct regulatory states in bidirectional promoters. *Epigenetics & Chromatin*, 11(1):66, 2018. URL https://doi.org/10.1186/s13072-018-0236-7

-1- Peter Ebert, Thomas Lengauer, and Christoph Bock. Epigenome-based prediction of gene expression across species. *bioRxiv*, 2018. URL https://doi.org/10.1101/371146

-1- Peter Ebert and Marcel H. Schulz. Fast Detection of Differential Chromatin Domains with SCIDDO. *bioRxiv*, 2018. URL https://doi.org/10.1101/441766

-n- Deborah Gerard, Florian Schmidt, Aurelien Ginolhac, Martine Schmitz, Rashi Halder, et al. Temporal epigenomic profiling identifies AHR as dynamic super-enhancer controlled regulator of mesenchymal multipotency. *bioRxiv*, page 183988, 2018. URL https://doi.org/10.1101/183988

-n- Abdulrahman Salhab, Karl Nordström, Gilles Gasparoni, Kathrin Kattler, Peter Ebert, et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biology*, 19(1):150, 2018. URL https://doi.org/10.1186/s13059-018-1510-5

-n- Florian Schmidt, Fabian Kern, Peter Ebert, Nina Baumgarten, and Marcel H Schulz. TEPIC 2 - An extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, 2018. URL https://doi.org/10.1093/bioinformatics/bty856

2017:

-2- Markus List, Peter Ebert, and Felipe Albrecht. Ten Simple Rules for Developing Usable Software in Computational Biology. *PLoS Computational Biology*, 13(1):e1005265, 2017. URL https://doi.org/10.1371/journal.pcbi.1005265

-n- Sarvesh Nikumbh, Peter Ebert, and Nico Pfeifer. All Fingers Are Not the Same: Handling Variable-Length Sequences in a Discriminative Setting Using Conformal Multi-Instance Kernels. In *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*, volume 88, pages 16:1—-16:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. URL https://doi.org/10.4230/LIPIcs.WABI.2017.16

-n- Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, 45(1):54–66, 2017. URL https://doi.org/10.1093/nar/gkw1061

-n- Christopher Schröder, Elsa Leitão, Stefan Wallner, Gerd Schmitz, Ludger Klein-Hitpass, et al. Regions of common inter-individual DNA methylation differences in human monocytes: genetic basis and potential function. *Epigenetics & Chromatin*, 10(1):37, 2017. URL https://doi.org/10.1186/s13072-017-0144-2

2016:

-n- Pawel Durek, Karl Nordström, Gilles Gasparoni, Abdulrahman Salhab, Christopher Kressler, et al. Epigenomic Profiling of Human CD4 + T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. *Immunity*, 45 (5):1148–1161, 2016. URL https://doi.org/10.1016/j.immuni.2016.10.022

-n- Matthias Farlik, Florian Halbritter, Fabian Müller, Fizzah A. Choudry, Peter Ebert, et al. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell*, 19(6):808–822, 2016. URL https://doi.org/10.1016/j.stem.2016.10.019

-n- Hendrik G. Stunnenberg, Sergio Abrignani, David Adams, Melanie de Almeida, Lucia Altucci, et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5):1145–1149, 2016. URL https://doi.org/10.1016/j.cell.2016.11.007

-n- Stefan Wallner, Christopher Schröder, Elsa Leitão, Tea Berulava, Claudia Haak, et al. Epigenetic dynamics of monocyte-to-macrophage differentiation. *Epigenetics and Chromatin*, 9(1):1–17, 2016. URL https://doi.org/10.1186/s13072-016-0079-z

2015:

-1- Peter Ebert, Fabian Müller, Karl Nordström, Thomas Lengauer, and Marcel H. Schulz. A general concept for consistent documentation of computational analyses. *Database*, 2015 (0):bav050–bav050, 2015. URL https://doi.org/10.1093/database/bav050

2014:

-n- Daniel Becker, Pavlo Lutsik, Peter Ebert, Christoph Bock, Thomas Lengauer, et al. BiQ Analyzer HiMod: An interactive software tool for high-throughput locus-specific analysis of 5-methylcytosine and its oxidized derivatives. *Nucleic Acids Research*, 42(May):501–507, 2014. URL https://doi.org/10.1093/nar/gku457

Other published work

- Peter Ebert and Christoph Bock. Improving reference epigenome catalogs by computational prediction. *Nature Biotechnology*, 33(4):354–355, 2015. URL https://doi.org/10.1038/nbt.3194
- Peter Ebert: *Analyzing Histone Tail Dynamics: A Normal Mode Based Approach*
  Master's Thesis, Saarland University 2011
  Supervisor: Prof. Dr. Andreas Hildebrandt
- Peter Ebert: *Computational Analyses of Genomic Human Papillomavirus Data*
  Bachelor's Thesis, Saarland University / Max Planck Institute for Informatics, 2009
  Supervisor: Prof. Dr. Thomas Lengauer, PhD

CHAPTER E

Appendix: License and Copyright Information

E.1 Manuscripts

If applicable, license and copyright information for material reused for Chapters 3, 4 and 5 is listed below.

E.1.1 Reproducibility in Computational Research

The manuscript Ebert et al. [64] was published in *Oxford Database*. The article was published under a Creative Commons license. This license grants the following rights:

> "This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
> You are not required to obtain permission to reuse this article."
> (See link "Permissions" in the online version of the article: doi.org/10.1093/database/bav050.)

Author contributions: none specified in the manuscript.
The author contributions for this project including the extensions presented in this thesis are as follows: all authors contributed to conceptualizing the published metadata specification. P.E. implemented the XSD process specification, the corresponding XML process templates, the CSS design, and the Python process/metadata validation script. Figures 1 and 3 in the publication were created by Fabian Müller; Figure 1 was reused in Chapter 3 (Figure 3.1). All versions of the "CHP" process were developed by P.E. and Andreas Richter (acknowledged in the publication) and implemented by P.E. The DEEP file tracking database including the Python SOAP client, the Python XML-RPC server, and the DEEP status website were conceptualized by P.E. with technical support by Joachim Büch and Georg Friedrich. The DEEP file tracking database was implemented by P.E. with assistance by Anna Hildebrandt (née Dehof). The Java file synchronization tool was implemented by Georg Friedrich. All implementation work for the Python SOAP client, the Python XML-RPC server, and the DEEP status website was carried out by P.E. with technical support by Joachim Büch and Georg Friedrich.

E.1.2 Fast Detection of Differential Chromatin Domains with SCIDDO

The manuscript Ebert and Schulz [63] is in preparation for a second submission (2018-11) and is publicly available as a preprint at bioRxiv: doi.org/10.1101/441766.

Author contributions as stated in the manuscript: "P.E. and M.H.S. conceptualized the project; P.E. carried out the implementation work, analyzed the data, and wrote the first draft of the manuscript; all authors contributed to the writing of the manuscript. All authors read and approved the final manuscript."

E.1.3 Epigenome-based Prediction of Gene Expression across Species

The manuscript Ebert et al. [65] has been submitted (submission: 2018-07; under revision at the time of writing: 2018-11) and is publicly available as a preprint at bioRxiv: doi.org/10.1101/371146.

Author contributions as stated in the manuscript: "P.E. and C.B. conceptualized the project with input from T.L.; P.E. analyzed the data and wrote the first draft of the manuscript; all authors contributed to the writing of the manuscript."

E.2 Figure Reprints

Table E.1: Licensing information for figure reprints

Figure	License	Publisher	Source
1.1	#4432540824975	Nature Publishing Group	Wagner [253]
2.1	#4407631152362	Nature Publishing Group	Felsenfeld and Groudine [77]
2.2	#4407621492444	Nature Publishing Group	Luger et al. [151]

Acronyms and Abbreviations

BAM binary alignment map 48, 61

BLUEPRINT European Hematopoietic Epigenome Project [1] 31, 58, 89, 108

bp base pairs 9, 11, 13, 18, 22, 23, 62, 73, 89

CGI CpG island 17, 85

ChIP-seq chromatin immunoprecipitation followed by high-throughput sequencing 19–25, 32, 50, 54, 61, 69, 85, 108, 109, 111

CSS Cascading Style Sheets 39, 42, 136

CSV comma-separated value 48

CWL Common Workflow Language 34

DAC Data Analysis Center 31, 37–39, 44, 47, 48, 50

DCC Data Collection Center 37, 39, 47, 48, 50, 56

DCD differential chromatin domain 60, 63, 69–75, 77–83, 120

DEEP Deutsches Epigenom Programm [48] 5, 18–20, 25, 27, 31, 32, 35–40, 42, 43, 45, 47, 48, 50–53, 55, 56, 58, 60, 61, 89, 94, 96, 107, 108, 111, 136

DEG differentially expressed gene 60, 62, 63, 74–79, 81–83

DNA deoxyribonucleic acid 8–13, 15–20, 23, 59, 85, 100, 103, 104

ENA European Nucleotide Archive 89, 98, 128

ENCODE Encyclopedia of DNA Elements [49, 231, 241] 31, 58, 63, 89, 94, 96, 108

FAANG Functional Annotation of Animal Genomes 35

Gbp giga base pairs 103

HBDE histone-binding domain enzyme 14, 17

HDAC histone deacetylase 14

HME histone-modifying enzyme 14, 16, 17

HTS high-throughput sequencing 3, 18, 19

IHEC International Human Epigenome Consortium [233] 14, 22, 32, 47, 50, 52, 58, 61, 107, 108

INSDC International Nucleotide Sequence Database Collaboration 55, 56

ISA-tab Investigation-Study-Assay tab-separated data 35, 36

IST Information Services and Technology 31, 44

JSON JavaScript Object Notation 33, 52, 111

KAT lysine acetyltransferase 14, 17

KDM lysine demethylase 14

KMT lysine methyltransferase 14

LUCA last universal common ancestor 2

MIBBI Minimum Reporting Guidelines for Biological and Biomedical Investigations 35

MINSEQE Minimum Information about a high-throughput Nucleotide Sequencing Experiment 35

MNase-seq micrococcal nuclease digestion followed by high-throughput sequencing 18

NGS next–generation sequencing 3, 5, 19

NIH National Institutes of Health 30

nm nanometer 10

NOMe-seq nucleosome occupancy and methylome sequencing 18

qPCR quantitative polymerase chain reaction 21

REMC NIH Roadmap Epigenomics Mapping Consortium 22, 31, 61, 69

RNA ribonucleic acid 8, 9

SCUFL2 Simple Conceptual Unified Flow Language (v2) 34

SRA Sequence Read Archive 89, 98, 128

TF transcription factor 12, 13, 15, 17, 19, 21, 85, 87

TFBS transcription factor binding site 85

TSS transcription start site 11, 89

TSV tab–separated value 48

UUID universally unique identifier 111

XML Extensible Markup Language 34, 39, 40, 42, 52, 136

XML-RPC Extensible Markup Language Remote Procedure Call 42, 136

XSD XML Schema Definition 39, 52, 136

Bibliography

[1] D Adams, L Altucci, Se Antonarakis, J Ballesteros, S Beck, et al. BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3), 2012. URL https://doi.org/10.1038/nbt.2153.

[2] Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1):W3–W10, 2016. URL https://doi.org/10.1093/nar/gkw343.

[3] Vikram Agarwal, George W. Bell, Jin-Wu Nam, and David P. Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4(AUGUST2015):1–38, 2015. URL https://doi.org/10.7554/eLife.05005.

[4] Bruce Alberts, Marc W Kirschner, Shirley Tilghman, and Harold Varmus. Rescuing US biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences*, 111(16):5773–5777, 2014. URL https://doi.org/10.1073/pnas.1404402111.

[5] Felipe Albrecht, Markus List, Christoph Bock, and Thomas Lengauer. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. *Nucleic Acids Research*, 44(W1):W581–W586, 2016. URL https://doi.org/10.1093/nar/gkw211.

[6] C. David Allis, Shelley L. Berger, Jacques Cote, Sharon Dent, Thomas Jenuwien, et al. New nomenclature for chromatin-modifying enzymes. *Cell*, 131(4):633–636, 2007. URL https://doi.org/10.1016/j.cell.2007.10.039.

[7] Geneviève Almouzni, Lucia Altucci, Bruno Amati, Neil Ashley, David Baulcombe, et al. Relationship between genome and epigenome - challenges and requirements for future research. *BMC Genomics*, 15(1):487, 2014. URL https://doi.org/10.1186/1471-2164-15-487.

[8] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. URL https://doi.org/10.1016/S0022-2836(05)80360-2.

[9] Robin Andersson, Stefan Enroth, Alvaro Rada-Iglesias, Claes Wadelius, and Jan Komorowski. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome research*, 19(10):1732–41, 2009. URL https://doi.org/10.1101/gr.092353.109.

[10] Laura Arrigoni, Andreas S. Richter, Emily Betancourt, Kerstin Bruder, Sarah Diehl, et al. Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Research*, 44(7):e67–e67, 2016. URL https://doi.org/10.1093/nar/gkv1495.

[11] Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology*, 9(11):e1003326, 2013. URL https://doi.org/10.1371/journal.pcbi.1003326.

[12] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Young Roh, Dustin E. Schones, et al. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007. URL https://doi.org/10.1016/j.cell.2007.05.009.

[13] Daniel Becker, Pavlo Lutsik, Peter Ebert, Christoph Bock, Thomas Lengauer, et al. BiQ Analyzer HiMod: An interactive software tool for high-throughput locus-specific analysis of 5-methylcytosine and its oxidized derivatives. *Nucleic Acids Research*, 42(May):501–507, 2014. URL https://doi.org/10.1093/nar/gku457.

[14] C. Glenn Begley and Lee M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391): 531–533, 2012. URL https://doi.org/10.1038/483531a.

[15] Fatemeh Behjati Ardakani, Kathrin Kattler, Karl Nordström, Nina Gasparoni, Gilles Gasparoni, et al. Integrative analysis of single-cell expression data reveals distinct regulatory states in bidirectional promoters. *Epigenetics & Chromatin*, 11(1):66, 2018. URL https://doi.org/10.1186/s13072-018-0236-7.

[16] Rimma Belotserkovskaya, Sangtaek Oh, Vladimir A Bondarenko, Vasily M Studitsky, Danny Reinberg, et al. FACT facilitates transription-dependent nucleosomes alteration. *Science*, 301(August):1090–1093, 2003.

[17] Rimma Belotserkovskaya, Abbie Saunders, John T. Lis, and Danny Reinberg. Transcription through chromatin: Understanding a complex FACT. *Biochimica et Biophysica Acta - Gene Structure and Expression*, 1677(1-3):87–99, 2004. URL https://doi.org/10.1016/j.bbaexp.2003.09.017.

[18] Bérénice A Benayoun, Elizabeth A Pollina, Duygu Ucar, Salah Mahmoudi, Kalpana Karra, et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell*, 158(3):673–88, 2014. URL https://doi.org/10.1016/j.cell.2014.06.027.

[19] Rinze Benedictus, Frank Miedema, and Mark W. J. Ferguson. Fewer numbers, better science. *Nature*, 538(7626):453–455, 2016. URL https://doi.org/10.1038/538453a.

[20] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. URL https://doi.org/10.2307/2346101.

[21] Dan Benveniste, Hans-Joachim Sonntag, Guido Sanguinetti, and Duncan Sproul. Transcription factor binding predicts histone modifications in human cell lines. *Proc Natl Acad Sci U S A*, 111(37):13367–13372, 2014. URL https://doi.org/10.1073/pnas.1412081111.

[22] Bradley E. Bernstein, Tarjei S. Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J. Huebert, et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326, 2006. URL https://doi.org/10.1016/j.cell.2006.02.041.

[23] Bradley E. Bernstein, Alexander Meissner, and Eric S. Lander. The mammalian epigenome. *Cell*, 128(4):669–81, 2007. URL https://doi.org/10.1016/j.cell.2007.01.033.

[24] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, et al. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, pages 319–326. Springer Berlin Heidelberg, 2008. URL https://doi.org/10.1007/978-3-540-78246-9_38.

[25] Ewan Birney, John A. Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R. Gingeras, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007. URL https://doi.org/10.1038/nature05874.

[26] J.M. Bland and D.G. Altman. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet*, 327:307–310, 1986. URL https://doi.org/10.1016/S0140-6736(86)90837-8.

[27] Christoph Bock. Epigenetic biomarker development. *Epigenomics*, 1(1):99–110, 2009. URL https://doi.org/10.2217/epi.09.6.

[28] Dario Boffelli, Marcelo A. Nobrega, and Edward M. Rubin. Comparative genomics at the vertebrate extremes. *Nature Reviews Genetics*, 5(6):456–465, 2004. URL https://doi.org/10.1038/nrg1350.

[29] Roberto Bonasio, Shengjiang Tu, and Danny Reinberg. Molecular signals of epigenetic states. *Science (New York, N.Y.)*, 330 (6004):612–6, 2010. URL https://doi.org/10.1126/science.1191078.

[30] James Bonner, Michael E. Dahmus, Douglas Fambrough, R.-c. C. Huang, Keiji Marushige, et al. The Biology of Isolated Chromatin: Chromosomes, biologically active in the test tube, provide a powerful tool for the study of gene action. *Science*, 159(3810):47–56, 1968. URL https://doi.org/10.1126/science.159.3810.47.

[31] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, 2011. URL https://doi.org/10.1038/nature10532.

[32] Alessandra Breschi, Thomas R. Gingeras, and Roderic Guigó. Comparative transcriptomics in human and mouse. *Nature Reviews Genetics*, 18(7):425–440, 2017. URL https://doi.org/10.1038/nrg.2017.19.

[33] Julie Brind'Amour, Sheng Liu, Matthew Hudson, Carol Chen, Mohammad M. Karimi, et al. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nature Communications*, 6:6033, 2015. URL https://doi.org/10.1038/ncomms7033.

[34] David M Budden, Daniel G Hurley, Joseph Cursons, John F Markham, Melissa J Davis, et al. Predicting expression: the complementary power of histone modification and transcription factor binding data. *Epigenetics & chromatin*, 7(1):36, 2014. URL https://doi.org/10.1186/1756-8935-7-36.

[35] Christa Buecker and Joanna Wysocka. Enhancers as information integration hubs in development: lessons from genomics. *Trends in Genetics*, 28(6):276–284, 2012. URL https://doi.org/10.1016/j.tig.2012.02.008.

[36] David Bujold, David Anderson de Lima Morais, Carol Gauthier, Catherine Côté, Maxime Caron, et al. The International Human Epigenome Consortium Data Portal. *Cell Systems*, 3(5):496–499.e2, 2016. URL https://doi.org/10.1016/j.cels.2016.10.019.

[37] Eliezer Calo and Joanna Wysocka. Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell*, 49(5): 825–837, 2013. URL https://doi.org/10.1016/j.molcel.2013.01.038.

[38] Eric I. Campos, James M. Stafford, and Danny Reinberg. Epigenetic inheritance: histone bookmarks across generations. *Trends in Cell Biology*, 24(11):664–674, 2014. URL https://doi.org/10.1016/j.tcb.2014.08.004.

[39] Enrique Carrillo-de Santa-Pau, David Juan, Vera Pancaldi, Felipe Were, Ignacio Martin-Subero, et al. Automatic identification of informative regions with epigenomic changes associated to hematopoiesis. *Nucleic Acids Research*, 45(16):9244–9259, 2017. URL https://doi.org/10.1093/nar/gkx618.

[40] Nilanjana Chatterjee, Justin A North, Mekonnen Lemma Dechassa, Mridula Manohar, Rashmi Prasad, et al. Histone Acetylation near the Nucleosome Dyad Axis Enhances Nucleosome Disassembly by RSC and SWI/SNF. *Molecular and Cellular Biology*, 35(23):4083–4092, 2015. URL https://doi.org/10.1128/MCB.00441-15.

[41] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, 2013. URL https://doi.org/10.1186/1471-2105-14-128.

[42] Yiwen Chen, Nicolas Negre, Qunhua Li, Joanna O Mieczkowska, Matthew Slattery, et al. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat Methods*, 9(6):609–614, 2012. URL https://doi.org/10.1038/nmeth.1985.

[43] Ho Ryun Chung, Ilona Dunkel, Franziska Heise, Christian Linke, Sylvia Krobitsch, et al. The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS ONE*, 5(12), 2010. URL https://doi.org/10.1371/journal.pone.0015754.

[44] Cedric R. Clapier, Janet Iwasa, Bradley R. Cairns, and Craig L. Peterson. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology*, 18(7):407–422, 2017. URL https://doi.org/10.1038/nrm.2017.26.

[45] Stephen J. Clark, Heather J. Lee, Sébastien A. Smallwood, Gavin Kelsey, and Wolf Reik. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biology*, 17(1):72, 2016. URL https://doi.org/10.1186/s13059-016-0944-x.

[46] David Clever, Rahul Roychoudhuri, Michael G. Constantinides, Michael H. Askenase, Madhusudhanan Sukumar, et al. Oxygen Sensing by T Cells Establishes an Immunologically Tolerant Metastatic Niche. *Cell*, 166(5):1117–1131.e14, 2016. URL https://doi.org/10.1016/j.cell.2016.07.032.

[47] Francis S. Collins and Lawrence A. Tabak. Policy: NIH plans to enhance reproducibility. *Nature*, 505(7485):612–613, 2014. URL https://doi.org/10.1038/505612a.

[48] DEEP Consortium. Deutsches Epigenom Programm, 2012. URL www.deutsches-epigenom-programm.de.

[49] E N C O D E Project Consortium, Richard M Myers, John Stamatoyannopoulos, Michael Snyder, Ian Dunham, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*, 9(4):e1001046, 2011. URL https://doi.org/10.1371/journal.pbio.1001046.

[50] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431 (7011):931–945, 2004. URL https://doi.org/10.1038/nature03001.

[51] Antoine Coulon, Carson C. Chow, Robert H. Singer, and Daniel R. Larson. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nature Reviews Genetics*, 14(8):572–584, 2013. URL https://doi.org/10.1038/nrg3484.

[52] Menno P. Creyghton, Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50): 21931–21936, 2010. URL https://doi.org/10.1073/pnas.1016071107.

[53] Pol Cuscó and Guillaume J. Filion. Zerone: A ChIP-seq discretizer for multiple replicates with built-in quality control. *Bioinformatics*, 32(19):2896–2902, 2016. URL https://doi.org/10.1093/bioinformatics/btw336.

[54] Charles Darwin. *On the origin of species*. John Murray, 1859.

[55] Carrie Deans and Keith A. Maggert. What Do You Mean, "Epigenetic"? *Genetics*, 199(4):887–896, 2015. URL https://doi.org/10.1534/genetics.114.173492.

[56] Job Dekker, Andrew S. Belmont, Mitchell Guttman, Victor O. Leshyk, John T. Lis, et al. The 4D nucleome project. *Nature*, 549(7671):219–226, 2017. URL https://doi.org/10.1038/nature23884.

[57] Aaron Diaz, Kiyoub Park, DA Lim, and JS Song. Normalization, bias correction, and peak calling for ChIP-seq. *Stat Appl Genet Mol Biol*, 11(3), 2012. URL https://doi.org/10.1515/1544-6115.1750.Normalization.

[58] Xianjun Dong, Melissa C Greven, Anshul Kundaje, Sarah Djebali, James B Brown, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, 13(9):R53, 2012. URL https://doi.org/10.1186/gb-2012-13-9-r53.

[59] Dr. Chris Drummond. Replicability is not reproducibility: Nor is it good science. *Proceedings of the Evaluation Methods for Machine Learning Workshop 26th International Conference for Machine Learning*, pages 1–4, 2009.

[60] Pawel Durek, Karl Nordström, Gilles Gasparoni, Abdulrahman Salhab, Christopher Kressler, et al. Epigenomic Profiling of Human CD4 + T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. *Immunity*, 45(5):1148–1161, 2016. URL https://doi.org/10.1016/j.immuni.2016.10.022.

[61] Charles R. Ebersole, Jordan R. Axt, and Brian A. Nosek. Scientists' Reputations Are Based on Getting It Right, Not Being Right. *PLOS Biology*, 14(5):e1002460, 2016. URL https://doi.org/10.1371/journal.pbio.1002460.

[62] Peter Ebert and Christoph Bock. Improving reference epigenome catalogs by computational prediction. *Nature Biotechnology*, 33(4):354–355, 2015. URL https://doi.org/10.1038/nbt.3194.

[63] Peter Ebert and Marcel H. Schulz. Fast Detection of Differential Chromatin Domains with SCIDDO. *bioRxiv*, 2018. URL https://doi.org/10.1101/441766.

[64] Peter Ebert, Fabian Müller, Karl Nordström, Thomas Lengauer, and Marcel H. Schulz. A general concept for consistent documentation of computational analyses. *Database*, 2015(0):bav050–bav050, 2015. URL https://doi.org/10.1093/database/bav050.

[65] Peter Ebert, Thomas Lengauer, and Christoph Bock. Epigenome-based prediction of gene expression across species. *bioRxiv*, 2018. URL https://doi.org/10.1101/371146.

[66] Thea A Egelhofer, Aki Minoda, Sarit Klugman, Kyungjoon Lee, Paulina Kolasinska-Zwierz, et al. An assessment of histone-modification antibody quality. *Nature Structural & Molecular Biology*, 18(1):91–93, 2011. URL https://doi.org/10.1038/nsmb.1972.

[67] Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016. URL https://doi.org/10.1073/pnas.1602413113.

[68] Hans Ellegren. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, 29(1):51–63, 2014. URL https://doi.org/10.1016/j.tree.2013.09.008.

[69] Christine G. Elsik, Deepak R. Unni, Colin M. Diesh, Aditi Tayal, Marianne L. Emery, et al. Bovine Genome Database: new tools for gleaning function from the Bos taurus genome. *Nucleic Acids Research*, 44(D1):D834–D839, 2016. URL https://doi.org/10.1093/nar/gkv1077.

[70] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–6, 2012. URL https://doi.org/10.1038/nmeth.1906.

[71] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–376, 2015. URL https://doi.org/10.1038/nbt.3157.

[72] Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D Ward, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011. URL https://doi.org/10.1038/nature09906.

[73] Timothy M Errington, Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax, et al. An open investigation of the reproducibility of cancer biology research. *eLife*, 3(16):e04333, 2014. URL https://doi.org/10.7554/eLife.04333.

[74] Manel Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–298, 2007. URL https://doi.org/10.1038/nrg2005.

[75] Matthias Farlik, Nathan C. Sheffield, Angelo Nuzzo, Paul Datlinger, Andreas Schönegger, et al. Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*, 10(8):1386–1397, 2015. URL https://doi.org/10.1016/j.celrep.2015.02.001.

[76] Matthias Farlik, Florian Halbritter, Fabian Müller, Fizzah A. Choudry, Peter Ebert, et al. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell*, 19(6):808–822, 2016. URL https://doi.org/10.1016/j.stem.2016.10.019.

[77] Gary Felsenfeld and Mark Groudine. Controlling the double helix. *Nature*, 421(6921):448–453, 2003. URL https://doi.org/10.1038/nature01411.

[78] Simon Fishilevich, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, 2017(1):1665–1680, 2017. URL https://doi.org/10.1093/database/bax028.

[79] Nuno A. Fonseca, Johan Rung, Alvis Brazma, and John C. Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 2012. URL https://doi.org/10.1093/bioinformatics/bts605.

[80] Harry G. Frankfurt. *On Inequality*. Princeton University Press, 2015.

[81] Alexey A. Fushan, Anton A. Turanov, Sang-Goo Lee, Eun Bae Kim, Alexei V. Lobanov, et al. Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, 14(3):352–365, 2015. URL https://doi.org/10.1111/acel.12283.

[82] Daniel J Gaffney, Graham McVicker, Athma a Pai, Yvonne N Fondufe-Mittendorf, Noah Lewellen, et al. Controls of nucleosome positioning in the human genome. *PLoS genetics*, 8(11):e1003036, 2012. URL https://doi.org/10.1371/journal.pgen.1003036.

[83] Charles Gawad, Winston Koh, and Stephen R. Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, 2016. URL https://doi.org/10.1038/nrg.2015.16.

[84] Deborah Gerard, Florian Schmidt, Aurelien Ginolhac, Martine Schmitz, Rashi Halder, et al. Temporal epigenomic profiling identifies AHR as dynamic super-enhancer controlled regulator of mesenchymal multipotency. *bioRxiv*, page 183988, 2018. URL https://doi.org/10.1101/183988.

[85] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, et al. What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17(6):669–681, 2007. URL https://doi.org/10.1101/gr.6339607.

[86] Raffaele Giancarlo, Simona E. Rombo, and Filippo Utro. Epigenomic k-mer dictionaries: shedding light on how sequence composition influences in vivo nucleosome positioning. *Bioinformatics*, 31(18):2939–2946, 2015. URL https://doi.org/10.1093/bioinformatics/btv295.

[87] Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–5, 2005. URL https://doi.org/10.1101/gr.4086505.

[88] Gregor D Gilfillan, Timothy Hughes, Ying Sheng, Hanne S Hjorthaug, Tobias Straub, et al. Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics*, 13(1):645, 2012. URL https://doi.org/10.1186/1471-2164-13-645.

[89] Aaron D. Goldberg, C. David Allis, and Emily Bernstein. Epigenetics: A Landscape Takes Shape. *Cell*, 128(4):635–638, 2007. URL https://doi.org/10.1016/j.cell.2007.02.006.

[90] L. Goodstadt. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics*, 26(21):2778–2779, 2010. URL https://doi.org/10.1093/bioinformatics/btq524.

[91] Vera Gorbunova, Andrei Seluanov, Zhengdong Zhang, Vadim N Gladyshev, and Jan Vijg. Comparative genetics of longevity and cancer: insights from long-lived rodents. *Nature reviews. Genetics*, 15(8):531–40, 2014. URL https://doi.org/10.1038/nrg3728.

[92] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, 2018. URL https://doi.org/10.1038/s41592-018-0046-7.

[93] Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, 2000. URL https://doi.org/10.1016/S0092-8674(00)81683-9.

[94] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, 2011. URL https://doi.org/10.1016/j.cell.2011.02.013.

[95] Bryan T Harada, William L Hwang, Sebastian Deindl, Nilanjana Chatterjee, Blaine Bartholomew, et al. Stepwise nucleosome translocation by RSC remodeling complexes. *eLife*, 5:e10051, 2016. URL https://doi.org/10.7554/eLife.10051.

[96] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9):1760–74, 2012. URL https://doi.org/10.1101/gr.135350.111.

[97] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009. URL https://doi.org/10.1007/b94608.

[98] Steven R Head, H Kiyomi Komori, Sarah A LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, et al. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques*, 56(2):61–4, 66, 68, passim, 2014. URL https://doi.org/10.2144/000114133.

[99] S. Blair Hedges, Julie Marin, Michael Suleski, Madeline Paymer, and Sudhir Kumar. Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution*, 32(4):835–845, 2015. URL https://doi.org/10.1093/molbev/msv037.

[100] Matthias Heinig, Maria Colomé-Tatché, Aaron Taudt, Carola Rintisch, Sebastian Schafer, et al. histoneHMM: Differential analysis of histone modifications with broad genomic footprints. *BMC Bioinformatics*, 16(1):60, 2015. URL https://doi.org/10.1186/s12859-015-0491-6.

[101] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–8, 2007. URL https://doi.org/10.1038/ng1966.

[102] Myriam Hemberger, Wendy Dean, and Wolf Reik. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nature Reviews Molecular Cell Biology*, 10(8):526–537, 2009. URL https://doi.org/10.1038/nrm2727.

[103] Andrew D. Higginson and Marcus R. Munafò. Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology*, 14(11):e2000995, 2016. URL https://doi.org/10.1371/journal.pbio.2000995.

[104] Joseph M. Hilbe. *Modeling Count Data*. Cambridge University Press, 2014. URL https://doi.org/10.1017/CBO9781139236065.

[105] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, 2012. URL https://doi.org/10.1038/nmeth.1937.

[106] Thomas J. Hudson, Warwick Anderson, Axel Aretz, Anna D. Barker, Cindy Bell, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010. URL https://doi.org/10.1038/nature08987.

[107] John P A Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):0696–0701, 2005. URL https://doi.org/10.1371/journal.pmed.0020124.

[108] John P. A. Ioannidis. How to Make More Published Research True. *PLoS Medicine*, 11(10):e1001747, 2014. URL https://doi.org/10.1371/journal.pmed.1001747.

[109] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, 343(6172):776–779, 2014. URL https://doi.org/10.1126/science.1247651.

[110] T Jenuwein. Translating the Histone Code. *Science*, 293(5532):1074–1080, 2001. URL https://doi.org/10.1126/science.1063127.

[111] Selin Jessa and Claudia L Kleinman. Chromswitch: a Flexible Method To Detect Chromatin State Switches. *Bioinformatics*, 34(February):2286–2288, 2018. URL https://doi.org/10.1093/bioinformatics/bty075.

[112] Ning Jiang, Lin Wang, Jing Chen, Luwen Wang, Lindsey Leach, et al. Conserved and divergent patterns of DNA methylation in higher vertebrates. *Genome biology and evolution*, 6(11):2998–3014, 2014. URL https://doi.org/10.1093/gbe/evu238.

[113] Peter A Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492, 2012. URL https://doi.org/10.1038/nrg3230.

[114] Peter A. Jones and Stephen B. Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, 2007. URL https://doi.org/10.1016/j.cell.2007.01.029.

[115] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahovicek, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2926–31, 2010. URL https://doi.org/10.1073/pnas.0909344107.

[116] S Karlin and S F Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences*, 90(12):5873–5877, 1993. URL https://doi.org/10.1073/pnas.90.12.5873.

[117] Samuel Karlin and Stephen F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–2268, 1990.

[118] Samuel Karlin, Amir Dembo, and Tsutomu Kawabata. Statistical Composition of High-Scoring Segments from Molecular Sequences. *The Annals of Statistics*, 18(2):571–581, 1990. URL https://doi.org/10.1214/aos/1176347616.

[119] M. Kasowski, S. Kyriazopoulou-Panagiotopoulou, F. Grubert, J. B. Zaugg, A. Kundaje, et al. Extensive Variation in Chromatin States Across Humans. *Science*, 342(6159):750–752, 2013. URL https://doi.org/10.1126/science.1242510.

[120] Michael Keane, Jeremy Semeiks, Andrew E. E. Webb, Yang I. I. Li, Víctor V??ctor V??ctor Quesada, et al. Insights into the evolution of longevity from the bowhead whale genome. *Cell Reports*, 10(1):112–122, 2015. URL https://doi.org/10.1016/j.celrep.2014.12.008.

[121] TK Kelly, Yaping Liu, FD Lay, and Gangning Liang. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*, pages 2497–2506, 2012. URL https://doi.org/10.1101/gr.143008.112.To.

[122] W James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11484–9, 2003. URL https://doi.org/10.1073/pnas.1932072100.

[123] Satyajeet P Khare, Farhat Habib, Rahul Sharma, Nikhil Gadewal, Sanjay Gupta, et al. HIstome–a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res*, 40(Database issue):D337–D342, 2012. URL https://doi.org/10.1093/nar/gkr1125.

[124] Benjamin L Kidder, Gangqing Hu, and Keji Zhao. ChIP-Seq: technical considerations for obtaining high-quality data. *Nature immunology*, 12(10):918–22, 2011. URL https://doi.org/10.1038/ni.2117.

[125] Tae Hoon Kim and Bing Ren. An all-round view of eukaryotic transcription. *Genome biology*, 7(7):323, 2006. URL https://doi.org/10.1186/gb-2006-7-7-323.

[126] Sarah Kinkley, Johannes Helmuth, Julia K. Polansky, Ilona Dunkel, Gilles Gasparoni, et al. reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4+ memory T cells. *Nature Communications*, 7:12514, 2016. URL https://doi.org/10.1038/ncomms12514.

[127] D. E. Knuth. Literate Programming. *The Computer Journal*, 27(2):97–111, 1984. URL https://doi.org/10.1093/comjnl/27.2.97.

[128] Klaus-Peter Koepfli, Benedict Paten, and Stephen J. O'Brien. The Genome 10K Project: A Way Forward. *Annual Review of Animal Biosciences*, 3(1):57–111, 2015. URL https://doi.org/10.1146/annurev-animal-090414-014900.

[129] Eugene V Koonin. Are There Laws of Genome Evolution? *PLoS Computational Biology*, 7(8):e1002173, 2011. URL https://doi.org/10.1371/journal.pcbi.1002173.

[130] Eugene V. Koonin and Artem S Novozhilov. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life*, 61 (2):99–111, 2009. URL https://doi.org/10.1002/iub.146.

[131] Johannes Köster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012. URL https://doi.org/10.1093/bioinformatics/bts480.

[132] Tony Kouzarides. Chromatin Modifications and Their Function. *Cell*, 128(4):693–705, 2007. URL https://doi.org/10.1016/j.cell.2007.02.005.

[133] Evgenia V Kriventseva, Fredrik Tegenfeldt, Tom J Petty, Robert M Waterhouse, Felipe A. Simão, et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, 43(D1):D250–D256, 2015. URL https://doi.org/10.1093/nar/gku1220.

[134] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016. URL https://doi.org/10.1093/nar/gkw377.

[135] Julia Ladewig, Philipp Koch, and Oliver Brüstle. Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nature Reviews Molecular Cell Biology*, 14(4):225–236, 2013. URL https://doi.org/10.1038/nrm3543.

[136] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. URL https://doi.org/10.1038/35057062.

[137] Story C Landis, Susan G Amara, Khusru Asadullah, Chris P Austin, Robi Blumenstein, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*, 490(7419):187–191, 2012. URL https://doi.org/10.1038/nature11556.

[138] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–31, 2012. URL https://doi.org/10.1101/gr.136184.111.

[139] Ben Lehner. Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14(3): 168–178, 2013. URL https://doi.org/10.1038/nrg3404.

[140] Bing Li, Michael Carey, and Jerry L. Workman. The Role of Chromatin during Transcription. *Cell*, 128(4):707–719, 2007. URL https://doi.org/10.1016/j.cell.2007.01.015.

[141] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–8, 2008. URL https://doi.org/10.1101/gr.078212.108.

[142] Benjamin J. Liebeskind, Claire D. McWhite, and Edward M. Marcotte. Towards Consensus Gene Ages. *Genome Biology and Evolution*, 8(6):1812–1823, 2016. URL https://doi.org/10.1093/gbe/evw113.

[143] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. URL https://doi.org/10.1109/18.61115.

[144] Shin Lin, Yiing Lin, Joseph R. Nery, Mark a. Urich, Alessandra Breschi, et al. Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 111(48):17224–17229, 2014. URL https://doi.org/10.1073/pnas.1413624111.

[145] Markus List, Peter Ebert, and Felipe Albrecht. Ten Simple Rules for Developing Usable Software in Computational Biology. *PLoS Computational Biology*, 13(1):e1005265, 2017. URL https://doi.org/10.1371/journal.pcbi.1005265.

[146] Hannah K Long, David Sims, Andreas Heger, Neil P Blackledge, Claudia Kutter, et al. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife*, 2:e00348, 2013. URL https://doi.org/10.7554/eLife.00348.

[147] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. URL https://doi.org/10.1186/s13059-014-0550-8.

[148] P T Lowary and J Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology*, 276(1):19–42, 1998. URL https://doi.org/10.1006/jmbi.1997.1494.

[149] Rebecca F. Lowdon, Hyo Sik Jang, and Ting Wang. Evolution of Epigenetic Regulation in Vertebrate Genomes. *Trends in Genetics*, 32(5):269–283, 2016. URL https://doi.org/10.1016/j.tig.2016.03.001.

[150] W. R. Luebben, N. Sharma, and J. K. Nyborg. Nucleosome eviction and activated transcription require p300 acetylation of histone H3 lysine 14. *Proceedings of the National Academy of Sciences*, 107(45):19254–19259, 2010. URL https://doi.org/10.1073/pnas.1009650107.

[151] Karolin Luger, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997. URL https://doi.org/10.1038/38444.

[152] Malcolm R Macleod, Susan Michie, Ian Roberts, Ulrich Dirnagl, Iain Chalmers, et al. Biomedical research: increasing value, reducing waste. *The Lancet*, 383(9912):101–104, 2014. URL https://doi.org/10.1016/S0140-6736(13)62329-6.

[153] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, 2015. URL https://doi.org/10.1016/j.cell.2015.05.002.

[154] Alessandro Mammana and Ho-Ryun Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*, 16(1):151, 2015. URL https://doi.org/10.1186/s13059-015-0708-z.

[155] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. URL https://doi.org/10.1214/aoms/1177730491.

[156] Raphaël Margueron and Danny Reinberg. Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics*, 11(4):285–296, 2010. URL https://doi.org/10.1038/nrg2752.

[157] Tobias Marschall, Manja Marz, Thomas Abeel, Louis Dijkstra, Bas E. Dutilh, et al. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):bbw089, 2016. URL https://doi.org/10.1093/bib/bbw089.

[158] Francesca Mattiroli, Yajie Gu, Tejas Yadav, Jeremy L. Balsbaugh, Michael R. Harris, et al. DNA-mediated association of two histone-bound complexes of yeast Chromatin Assembly Factor-1 (CAF-1) drives tetrasome assembly in the wake of DNA replication. *eLife*, 6:1–23, 2017. URL https://doi.org/10.7554/eLife.22799.

[159] Ian Maze, Kyung-Min Noh, Alexey A. Soshnev, and C. David Allis. Every amino acid matters: essential contributions of histone variants to mammalian development and disease. *Nature Reviews Genetics*, 15(4):259–271, 2014. URL https://doi.org/10.1038/nrg3673.

[160] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[161] Jennifer R. S. Meadows and Kerstin Lindblad-Toh. Dissecting evolution and disease using comparative vertebrate genomics. *Nature Reviews Genetics*, 18(10):624–636, 2017. URL https://doi.org/10.1038/nrg.2017.51.

[162] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science*, 338(6114):1593–1599, 2012. URL https://doi.org/10.1126/science.1228186.

[163] Michael L Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, 2010. URL https://doi.org/10.1038/nrg2626.

[164] Mariann Micsinai, Fabio Parisi, Francesco Strino, Patrik Asp, Brian D Dynlacht, et al. Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic acids research*, 40(9):e70, 2012. URL https://doi.org/10.1093/nar/gks048.

[165] Tarjei S. Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007. URL https://doi.org/10.1038/nature06008.

[166] K. Jarrod Millman and Michael Aivazis. Python for Scientists and Engineers. *Computing in Science & Engineering*, 13(2): 9–12, 2011. URL https://doi.org/10.1109/MCSE.2011.36.

[167] Tanmoy Mondal, Markus Rasmussen, G. K. Pandey, A. Isaksson, and C. Kanduri. Characterization of the RNA content of chromatin. *Genome Research*, 20(7):899–907, 2010. URL https://doi.org/10.1101/gr.103473.109.

[168] Stefanie Mühlhausen, Hans Dieter Schmitt, Kuan-Ting Pan, Uwe Plessmann, Henning Urlaub, et al. Endogenous Stochastic Decoding of the CUG Codon by Competing Ser- and Leu-tRNAs in Ascoidea asiatica. *Current Biology*, 28(13):2046–2057.e5, 2018. URL https://doi.org/10.1016/j.cub.2018.04.085.

[169] Asher Mullard. Reliability of ’new drug target’ claims called into question. *Nature Reviews Drug Discovery*, 10(9):643–644, 2011. URL https://doi.org/10.1038/nrd3545.

[170] Ryuichiro Nakato and Katsuhiko Shirahige. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in Bioinformatics*, page bbw023, 2016. URL https://doi.org/10.1093/bib/bbw023.

[171] Nature. Journals unite for reproducibility. *Nature*, 515(7525):7–7, 2014. URL https://doi.org/10.1038/515007a.

[172] Nature. Reality check on reproducibility. *Nature*, 533(7604):437–437, 2016. URL https://doi.org/10.1038/533437a.

[173] Yoshihito Niimura and Masatoshi Nei. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene*, 346:13–21, 2005. URL https://doi.org/10.1016/j.gene.2004.09.025.

[174] Sarvesh Nikumbh, Peter Ebert, and Nico Pfeifer. All Fingers Are Not the Same: Handling Variable-Length Sequences in a Discriminative Setting Using Conformal Multi-Instance Kernels. In *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*, volume 88, pages 16:1—-16:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. URL https://doi.org/10.4230/LIPIcs.WABI.2017.16.

[175] Haruko Obokata, Teruhiko Wakayama, Yoshiki Sasai, Koji Kojima, Martin P. Vacanti, et al. Retraction: Stimulus-triggered fate conversion of somatic cells into pluripotency. *Nature*, 511(7507):112–112, 2014. URL https://doi.org/10.1038/nature13598.

[176] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004. URL https://doi.org/10.1093/bioinformatics/bth361.

[177] Travis E. Oliphant. Python for Scientific Computing. *Computing in Science & Engineering*, 9(3):10–20, 2007. URL https://doi.org/10.1109/MCSE.2007.58.

[178] George Orphanides and Danny Reinberg. A Unified Theory of Gene Expression. *Cell*, 108(4):439–451, 2002. URL https://doi.org/10.1016/S0092-8674(02)00655-4.

[179] Minhee Park, Albert J. Keung, and Ahmad S. Khalil. The epigenome: the next substrate for engineering. *Genome Biology*, 17 (1):183, 2016. URL https://doi.org/10.1186/s13059-016-1046-5.

[180] Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017. URL https://doi.org/10.1038/nmeth.4197.

[181] Neidhard Paweletz. Walther Flemming: pioneer of mitosis research. *Nature Reviews Molecular Cell Biology*, 2(1):72–75, 2001. URL https://doi.org/10.1038/35048077.

[182] William R Pearson. Selecting the Right Similarity-Scoring Matrix. In *Current Protocols in Bioinformatics*, volume 43, pages 3.5.1–3.5.9. John Wiley & Sons, Inc., 2013. URL https://doi.org/10.1002/0471250953.bi0305s43.

[183] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[184] R. D. Peng. Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227, 2011. URL https://doi.org/10.1126/science.1213847.

[185] Len A. Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4):288–295, 2013. URL https://doi.org/10.1038/nrg3458.

[186] Fernando Perez and Brian E. Granger. IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3):21–29, 2007. URL https://doi.org/10.1109/MCSE.2007.53.

[187] Svetlana Petruk, Yurii Sedkov, Danika M. Johnston, Jacob W. Hodgson, Kathryn L. Black, et al. TrxG and PcG Proteins but Not Methylated Histones Remain Associated with DNA through Replication. *Cell*, 150(5):922–933, 2012. URL https://doi.org/10.1016/j.cell.2012.06.046.

[188] Jennifer L. Plank and Ann Dean. Enhancer Function: Mechanistic and Genome-Wide Insights Come Together. *Molecular Cell*, 55(1):5–14, 2014. URL https://doi.org/10.1016/j.molcel.2014.06.015.

[189] Konstantin Y. Popadin, Maria Gutierrez-Arcelus, Tuuli Lappalainen, Alfonso Buil, Julia Steinberg, et al. Gene Age Predicts the Strength of Purifying Selection Acting on Gene Expression Variation in Humans. *The American Journal of Human Genetics*, 95(6):660–674, 2014. URL https://doi.org/10.1016/j.ajhg.2014.11.003.

[190] Karl Raimund Popper. *The Logic of Scientific Discovery*. Taylor & Francis Group, 1959.

[191] Madapura M Pradeepa, Graeme R Grimes, Yatendra Kumar, Gabrielle Olley, Gillian C A Taylor, et al. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature Genetics*, 48(6):681–686, 2016. URL https://doi.org/10.1038/ng.3550.

[192] Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–712, 2011. URL https://doi.org/10.1038/nrd3439-c1.

[193] M. Ptashne. Epigenetics: Core misconcept. *Proceedings of the National Academy of Sciences*, 110(18):7101–7103, 2013. URL https://doi.org/10.1073/pnas.1305399110.

[194] Mark Ptashne. On the use of the word 'epigenetic'. *Current Biology*, 17(7):R233–6, 2007. URL https://doi.org/10.1016/j.cub.2007.02.030.

[195] Sangya Pundir, Maria J. Martin, and Claire O'Donovan. UniProt Tools. *Current Protocols in Bioinformatics*, 53:1.29.1–1.29.15, 2016. URL https://doi.org/10.1002/0471250953.bi0129s53.

[196] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26 (6):841–842, 2010. URL https://doi.org/10.1093/bioinformatics/btq033.

[197] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283, 2011. URL https://doi.org/10.1038/nature09692.

[198] Vardhman K. Rakyan, Thomas A. Down, David J. Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541, 2011. URL https://doi.org/10.1038/nrg3000.

[199] Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn a Grüning, and Thomas Manke. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, 42(Web Server issue):187–91, 2014. URL https://doi.org/10.1093/nar/gku365.

[200] Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, 2016. URL https://doi.org/10.1093/nar/gkw257.

[201] Naim U. Rashid, Paul G. Giresi, Joseph G. Ibrahim, Wei Sun, and Jason D. Lieb. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*, 12(7):R67, 2011. URL https://doi.org/10.1186/gb-2011-12-7-r67.

[202] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, et al. The Human Cell Atlas. *eLife*, 6, 2017. URL https://doi.org/10.7554/eLife.27041.

[203] Ho Sung Rhee and B. Franklin Pugh. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, 147(6):1408–1419, 2011. URL https://doi.org/10.1016/j.cell.2011.11.013.

[204] Maria Aurelia Ricci, Carlo Manzo, María Filomena García-Parajo, Melike Lakadamyali, and Maria Pia Cosma. Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo. *Cell*, 160(6):1145–1158, 2015. URL https://doi.org/10.1016/j.cell.2015.01.054.

[205] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. URL https://doi.org/10.1038/nature14248.

[206] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172, 2015. URL https://doi.org/10.1038/nbt.3383.

[207] P. J. Rugg-Gunn, B. J. Cox, A. Ralston, and J. Rossant. Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proceedings of the National Academy of Sciences*, 107(24):10783–10790, 2010. URL https://doi.org/10.1073/pnas.0914507107.

[208] W L Ruzzo and M Tompa. A linear time algorithm for finding all maximal scoring subsequences. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, pages 234–41, 1999.

[209] Anjanabha Saha, Jacqueline Wittmeyer, and Bradley R. Cairns. Chromatin remodelling: The industrial revolution of DNA around histones. *Nature Reviews Molecular Cell Biology*, 7(6):437–447, 2006. URL https://doi.org/10.1038/nrm1945.

[210] Abdulrahman Salhab, Karl Nordström, Gilles Gasparoni, Kathrin Kattler, Peter Ebert, et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biology*, 19(1):150, 2018. URL https://doi.org/10.1186/s13059-018-1510-5.

[211] Steven L. Salzberg. Reminder to deposit DNA sequences. *Nature*, 533(7602):179–179, 2016. URL https://doi.org/10.1038/533179a.

[212] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):1–4, 2013. URL https://doi.org/10.1371/journal.pcbi.1003285.

[213] Christian Schmidl, André F Rendeiro, Nathan C Sheffield, and Christoph Bock. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods*, 12(10):963–965, 2015. URL https://doi.org/10.1038/nmeth.3542.

[214] Dominic Schmidt, Michael D Wilson, Benoit Ballester, Petra C Schwalie, Gordon D Brown, et al. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*, 328(5981):1036–1040, 2010. URL https://doi.org/10.1126/science.1186176.

[215] Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, 45(1): 54–66, 2017. URL https://doi.org/10.1093/nar/gkw1061.

[216] Florian Schmidt, Fabian Kern, Peter Ebert, Nina Baumgarten, and Marcel H Schulz. TEPIC 2 - An extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics*, 2018. URL https://doi.org/10.1093/bioinformatics/bty856.

[217] Santiago Schnell. Ten Simple Rules for a Computational Biologist's Laboratory Notebook. *PLoS computational biology*, 11 (9):e1004385, 2015. URL https://doi.org/10.1371/journal.pcbi.1004385.

[218] Dustin E. Schones and Keji Zhao. Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9 (3):179–191, 2008. URL https://doi.org/10.1038/nrg2270.

[219] Dustin E. Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, et al. Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, 132(5):887–898, 2008. URL https://doi.org/10.1016/j.cell.2008.02.022.

[220] Christopher Schröder, Elsa Leitão, Stefan Wallner, Gerd Schmitz, Ludger Klein-Hitpass, et al. Regions of common inter-individual DNA methylation differences in human monocytes: genetic basis and potential function. *Epigenetics & Chromatin*, 10(1):37, 2017. URL https://doi.org/10.1186/s13072-017-0144-2.

[221] Bernd Schuettengruber, Daniel Chourrout, Michel Vervoort, Benjamin Leblanc, and Giacomo Cavalli. Genome regulation by polycomb and trithorax proteins. *Cell*, 128(4):735–45, 2007. URL https://doi.org/10.1016/j.cell.2007.02.009.

[222] Omer Schwartzman and Amos Tanay. Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics*, 16(12):716–726, 2015. URL https://doi.org/10.1038/nrg3980.

[223] Eran Segal and Jonathan Widom. What controls nucleosome positions? *Trends in Genetics*, 25(8):335–43, 2009. URL https://doi.org/10.1016/j.tig.2009.06.002.

[224] Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, 450(7171):908–912, 2007. URL https://doi.org/10.1038/nature06397.

[225] Nathan C Sheffield and Christoph Bock. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics (Oxford, England)*, 32(4):587–9, 2016. URL https://doi.org/10.1093/bioinformatics/btv612.

[226] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–32, 2014. URL https://doi.org/10.1038/nrg3642.

[227] Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, 32(17):i639–i648, 2016. URL https://doi.org/10.1093/bioinformatics/btw427.

[228] Paul E. Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society Open Science*, 3(9):160384, 2016. URL https://doi.org/10.1098/rsos.160384.

[229] Charlotte Soneson, Michael I. Love, and Mark D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4(2):1521, 2016. URL https://doi.org/10.12688/f1000research.7563.2.

[230] Jimin Song and Kevin C Chen. Spectacle: fast chromatin state annotation using spectral learning. *Genome biology*, 16(1):33, 2015. URL https://doi.org/10.1186/s13059-015-0598-0.

[231] John A Stamatoyannopoulos, Michael Snyder, Ross Hardison, Bing Ren, Thomas Gingeras, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, 13(8):418, 2012. URL https://doi.org/10.1186/gb-2012-13-8-418.

[232] John D Storey, Andrew J Bass, Alan Dabney, and David Robinson. qvalue: Q-value estimation for false discovery rate control, 2015. URL http://github.com/jdstorey/qvalue.

[233] Hendrik G. Stunnenberg, Sergio Abrignani, David Adams, Melanie de Almeida, Lucia Altucci, et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5):1145–1149, 2016. URL https://doi.org/10.1016/j.cell.2016.11.007.

[234] Paul B. Talbert and Steven Henikoff. Histone variants — ancient wrap artists of the epigenome. *Nature Reviews Molecular Cell Biology*, 11(4):264–275, 2010. URL https://doi.org/10.1038/nrm2861.

[235] Paul B Talbert, Kami Ahmad, Geneviève Almouzni, Juan Ausió, Frederic Berger, et al. A unified phylogeny-based nomenclature for histone variants. *Epigenetics & Chromatin*, 5(1):7, 2012. URL https://doi.org/10.1186/1756-8935-5-7.

[236] Li Tan, Zhonghe Ke, Gregory Tombline, Nicholas Macoretta, Kevin Hayes, et al. Naked Mole Rat Cells Have a Stable Epigenome that Resists iPSC Reprogramming. *Stem Cell Reports*, 9(5):1721–1734, 2017. URL https://doi.org/10.1016/j.stemcr.2017.10.001.

[237] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015. URL https://doi.org/10.1093/bioinformatics/btv098.

[238] Chris F Taylor, Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, 26(8):889–896, 2008. URL https://doi.org/10.1038/nbt.1411.

[239] Leonid Teytelman, Deborah M Thurtle, Jasper Rine, and Alexander van Oudenaarden. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 110(46):18602–7, 2013. URL https://doi.org/10.1073/pnas.1316064110.

[240] Douglas Thain, Peter Ivie, and Haiyan Meng. Techniques for Preserving Scientific Software Executions: Preserve the Mess or Encourage Cleanliness? In *12th International Conference on Digital Preservation (iPres)*, 2015.

[241] The ENCODE Project Consortium, Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. URL https://doi.org/10.1038/nature11247.

[242] Reuben Thomas, Sean Thomas, Alisha K. Holloway, and Katherine S. Pollard. Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics*, 18(3):441–450, 2017. URL https://doi.org/10.1093/bib/bbw035.

[243] Cole Trapnell and Steven L. Salzberg. How to map billions of short reads onto genomes. *Nature Biotechnology*, 27(5):455–457, 2009. URL https://doi.org/10.1038/nbt0509-455.

[244] Jay J. Van Bavel, Peter Mende-Siedlecki, William J. Brady, and Diego A. Reinero. Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113(23):6454–6459, 2016. URL https://doi.org/10.1073/pnas.1521897113.

[245] Thijn van der Heijden, Joke J F A van Vugt, Colin Logie, and John van Noort. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proceedings of the National Academy of Sciences of the United States of America*, 109(38):E2514–22, 2012. URL https://doi.org/10.1073/pnas.1205659109.

[246] Juan Manuel Vazquez, Michael Sulak, Sravanthi Chigurupati, and Vincent J. Lynch. A Zombie LIF Gene in Elephants Is Upregulated by TP53 to Induce Apoptosis in Response to DNA Damage. *Cell Reports*, 24(7):1765–1776, 2018. URL https://doi.org/10.1016/j.celrep.2018.07.042.

[247] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, et al. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, 2001. URL https://doi.org/10.1126/science.1058040.

[248] Diego Villar, Paul Flicek, and Duncan T Odom. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature Reviews Genetics*, 15(4):221–233, 2014. URL https://doi.org/10.1038/nrg3481.

[249] Diego Villar, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, et al. Enhancer Evolution across 20 Mammalian Species. *Cell*, 160(3):554–566, 2015. URL https://doi.org/10.1016/j.cell.2015.01.006.

[250] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–8, 2009. URL https://doi.org/10.1038/nature07730.

[251] Philipp Voigt, Gary LeRoy, William J. Drury, Barry M. Zee, Jinsook Son, et al. Asymmetrically Modified Nucleosomes. *Cell*, 151(1):181–193, 2012. URL https://doi.org/10.1016/j.cell.2012.09.002.

[252] Philipp Voigt, Wee-Wei Wei Tee, and Danny Reinberg. A double take on bivalent promoters. *Genes and Development*, 27(12):1318–1338, 2013. URL https://doi.org/10.1101/gad.219626.113.

[253] Günter P. Wagner. The developmental genetics of homology. *Nature Reviews Genetics*, 8(6):473–479, 2007. URL https://doi.org/10.1038/nrg2099.

[254] Stefan Wallner, Christopher Schröder, Elsa Leitão, Tea Berulava, Claudia Haak, et al. Epigenetic dynamics of monocyte-to-macrophage differentiation. *Epigenetics and Chromatin*, 9(1):1–17, 2016. URL https://doi.org/10.1186/s13072-016-0079-z.

[255] Zhibin Wang, Chongzhi Zang, Jeffrey A. Rosenfeld, Dustin E. Schones, Artem Barski, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, 2008. URL https://doi.org/10.1038/ng.154.

[256] Warren A Whyte, David A Orlando, Denes Hnisz, Brian J Abraham, Charles Y Lin, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–19, 2013. URL https://doi.org/10.1016/j.cell.2013.03.035.

[257] E. O. Wilson. *Consilience: The Unity of Knowledge*. Vintage Books, Random House Inc, New York, 1999.

[258] Michael D Wilson and Duncan T Odom. Evolution of transcriptional control in mammals. *Current opinion in genetics & development*, 19(6):579–85, 2009. URL https://doi.org/10.1016/j.gde.2009.10.003.

[259] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic acids research*, 41(Web Server issue):W557–61, 2013. URL https://doi.org/10.1093/nar/gkt328.

[260] Jerry L. Workman. Nucleosome displacement in transcription. *Genes & Development*, 20(15):2009–2017, 2006. URL https://doi.org/10.1101/gad.1435706.

[261] Shu Xiao, Dan Xie, Xiaoyi Cao, Pengfei Yu, Xiaoyun Xing, et al. Comparative Epigenomic Annotation of Regulatory DNA. *Cell*, 149(6):1381–1392, 2012. URL https://doi.org/10.1016/j.cell.2012.04.029.

[262] Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, 2015.

[263] Song Yang, Nir Oksenberg, Sachiko Takayama, Seok-Jin Heo, Alexander Poliakov, et al. Functionally conserved enhancers with divergent sequences in distant vertebrates. *BMC Genomics*, 16(1):882, 2015. URL https://doi.org/10.1186/s12864-015-2070-7.

[264] Angela Yen and Manolis Kellis. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nature communications*, 6:7973, 2015. URL https://doi.org/10.1038/ncomms8973.

[265] Gabriel E Zentner, Paul J Tesar, and Peter C Scacheri. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8):1273–83, 2011. URL https://doi.org/10.1101/gr.122382.111.

[266] Daniel R Zerbino, Steven P Wilder, Nathan Johnson, Thomas Juettemann, and Paul R Flicek. The Ensembl Regulatory Build. *Genome Biology*, 16(1):56, 2015. URL https://doi.org/10.1186/s13059-015-0621-5.

[267] Yanxiao Zhang, Y.-H. Lin, Timothy D Johnson, Laura S Rozek, and Maureen A Sartor. PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*, 30(18):2568–2575, 2014. URL https://doi.org/10.1093/bioinformatics/btu372.

[268] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137.1–9, 2008. URL https://doi.org/10.1186/gb-2008-9-9-r137.

[269] Wei Zhao, Xiaping He, Katherine A Hoadley, Joel S Parker, David Hayes, et al. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*, 15(1):419, 2014. URL https://doi.org/10.1186/1471-2164-15-419.

[270] Xiangqun Zheng-Bradley, Johan Rung, Helen Parkinson, and Alvis Brazma. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology*, 11(12):R124, 2010. URL https://doi.org/10.1186/gb-2010-11-12-r124.

[271] Hui Zhi, Xin Li, Peng Wang, Yue Gao, Baoqing Gao, et al. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research*, 46, 2017. URL https://doi.org/10.1093/nar/gkx1020.

[272] Mark Ziemann, Yotam Eren, and Assam El-Osta. Gene name errors are widespread in the scientific literature. *Genome Biology*, 17(1):177, 2016. URL https://doi.org/10.1186/s13059-016-1044-7.