Bioinformatic Approaches for Understanding Chromatin Regulation

Juliane Perner

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Berlin, Juni 2015

Erstgutachter: Prof. Dr. Martin Vingron
Zweitgutachter: Prof. Dr. Denis Thieffry

Tag der Disputation: 24. Juli 2015

Preface

Acknowledgements

I am very grateful to Martin Vingron for many different aspects of his supervision: For his support, ideas, advice and discussions throughout my thesis. For introducing me to the field of chromatin biology and providing me the freedom to choose my research direction. For initiating many fruitful collaborations and for his mentorship and career advice throughout the years. For giving me the chance to gain teaching experience by assisting in one of his classes. For encouraging me to attend many interesting conferences, summer schools, symposia or meetings and the resulting scientific discussions and knowledge I gained. And finally, for providing an enjoyable and stimulating research environment in his group.

I wish to thank Ho-Ryun Chung for his constant interest in my work, especially in the promoter-centered chromatin signaling network. Through the almost daily scientific discussions with him I learned a lot about science, chromatin biology and the role of histone modifications in regulating transcription. I am very grateful to him for sharing his expertise with me and for his important contribution to the project. I also thank him for helping me through many cycles of revision of the manuscript.

I am grateful to Alfonso Valencia for allowing me to stay in his lab for a scientific internship and for the scientific discussions and advice throughout my stay and after. I also wish to thank the people in the department of Structural Biology and BioComputing at the Centro Nacional de Investigaciones Oncológicas (CNIO) in Madrid, Spain for making my stay such an enjoyable time. In particular, I thank Daniel Rico for helping me organizing the stay, for coffees and lively conversations. I also thank Vera Pancaldi for sharing her expertise on biological networks and for taking me around in Madrid. Further, I am grateful to Daniel Rico, Enrique Carrillo de Santa Pau and David Juan for initializing the mouse project and the lively discussions in person and over phone, skype or googledocs.

I wish to acknowledge the members of the BLUEPRINT Consortium, especially the people participating in the annual Analysis Jamboree, for sharing their experience in analyzing ChIP-Seq, RNA-Seq or sequencing-based methylation data. Further, I would like to thank the BLUEPRINT Consortium for providing me funding for my scientific internship at the CNIO in Madrid, Spain.

Many side projects that I have been working on are not part of this thesis. Nevertheless, they contributed largely to my working experience, knowledge and skills. I would like to thank Shin Eh, Phillip Grote and Bernhard Herrmann for performing the experiments and

providing data on histone modifications and gene expression in mesoderm development. This project and the accompanying discussions largely contributed to my understanding of the role of histone modifications in transcriptional regulation. I thank Phillip Grote for many interesting discussion about the Polycomb complex and its role in gene regulation. I also thank my Bachelor student Adrian Bierling who worked on a follow-up project investigating the function of histone modifications during differentiation. I wish to thank Christine Sers with whom I worked on understanding the effect of RAS-mediated signaling on transcription. Finally, I would like to thank the DREAM-team consisting of Alena van Bömmel, Alessandro Mammana, Brian Caffrey, Matthias Heinig and Matt Huska for our joint effort on solving the DREAM8 Toxicogenetics challange.

I am very grateful to all present and past members of the department of Computational Molecular Biology for the inspiring and pleasant working atmosphere I enjoyed during my PhD. I would like to thank my office mates Mike Love, Ruping Sun, Ho-Ryun Chung, Stephanie Schöne, Alena van Bömmel and Xinyi Yang for scientific discussions, the cheerful atmosphere and a wonderful time at the lab. The present thesis improved a lot due to the proofreading and helpful comments of Alessandro Mammana, Anna Ramisch and Ho-Ryung Chung. I also wish to thank the International Max Planck Research School for Computational Biology, especially the coordinator Kirsten Kelleher and all fellow students that I have met within this program for science- and non-science-related discussion, friendships and many cheerful hours during my PhD.

Last but not least, I am deeply grateful to my parents who helped me through all troubles I encountered with their constant support, love and life experience.

Publications

The analysis presented in Chapter 6 has been published in the Journal *Nucleic Acids Research* [1]. I was involved in designing the study, implementing and performing the presented analysis, interpreting the results and writing the manuscript. I would like to acknowledge Ho-Ryun Chung for his invaluable contribution to this manuscript. He contributed largely to the design of the study, interpreting the results and developing the "Polycomb-hypothesis", as well as writing the manuscript. I would like to acknowledge Julia Lasserre for computing the Sparse Partial Correlation Network and for her discussions throughout the preparations of this manuscript. Further, I would like to acknowledge Sarah Kinkley for performing the Co-IP experiments to validate the novel interactions. I would also like to thank Ho-Ryun Chung for writing up ideas and discussions, which resulted from the work on the main manuscript, in [2]. I contributed to this manuscript by analyzing and interpreting the data.

A manuscript containing parts of the work presented in Chapter 7 is about to be submitted and a pre-print can be found in the bioRxiv online repository [3]. I was involved in designing the study, collecting the data, performing the co-localization analysis, interpreting the results, deriving the "communication-hypothesis" and writing the manuscript. With regards to Chapter 7, I would like to acknowledge the significant contribution of Enrique Carrillo de Santa Pau, David Juan, and Daniel Rico. They performed the preprocessing of the data, ran the ChromHMM and the enrichment analysis, and derived the sub-networks. Further, they contributed ideas and discussed the results presented in Chapter 7.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

What defines cellular identity? How are the morphology and the function of a cell maintained, while leaving flexibility to respond to environmental stimuli or to proceed through the cell cycle? The necessary information is stored in the genome: the heritable molecular entity that carries the complete building instructions for a whole organism. Every cell in a multicellular organism contains an identical copy of the same genome. Yet, they show striking differences in their characteristics. These differences are already manifested in the differential usage of the available functional units called genes.

This being the case, a key cellular process is transcription, which is the complex process of copying genes into RNA. The RNA molecules in turn function as blueprints for proteins, which are the building blocks of the cell, or act as functional units themselves. With such a profound role, it is clear that transcription has to be tightly regulated in time and space. However, despite being the focus of current intense research, our understanding of the various factors influencing the process and the regulation of transcription is still incomplete.

Accumulating evidence identified chromatin as a central component in transcriptional regulation. Chromatin refers to the packing of DNA together with proteins in the cell nucleus. The role of chromatin is two-fold. First, the strength of the packing defines the accessibility of the underlying genes. Second, the histone proteins, around which the DNA is wrapped for packing, can be chemically modified. These *histone modifications* are mediated and read by other proteins, referred to as *chromatin modifiers* in the remainder. The chromatin modifiers in turn can communicate the signals to the transcriptional machinery. This sequences of multiple specific interactions resemble signaling pathways playing a role in transcriptional regulation. However, the individual interactions in the chromatin signaling pathways are often unknown. Thus, identifying the interactions between the histone modifications and chromatin modifiers would provide a basis for understanding how the transcriptional status of a gene is regulated.

Systematic experimental endeavors allow for a global analysis of the interactions between histone modifications and chromatin modifiers. The experimental techniques behind these endeavors allow to map the precise DNA-binding locations of proteins along the whole genome. While histone modifications have been in the focus of many research projects, only recently the first systematic experiments detecting the genome-wide DNA-binding events of chromatin modifiers were produced (e.g. in [4]). The availability of large genome-wide DNA-protein binding data sets allows for the identification of co-localizing proteins along the whole genome using computational means. Under the assumption that co-localization shows common function, we can infer biological hypotheses on chromatin signaling. However, in a typical genome-wide study only pair-wise, hypothesis-driven comparisons are performed. This approach has the disadvantage that intermediate interaction partners are not taken into account. Without additional, often difficult experiments the pairwise comparisons might lead to the inference of indirect interactions. By integrating many genome-wide experiments and applying machine learning methods, we can account for all proteins for which data is available. The inferred interactions are more likely to correspond to true mechanistic interactions.

1.1. Research objective

The individual interactions between chromatin-related proteins can be reconstructed from genome-wide DNA-protein binding data by applying network reconstruction methods. The resulting network view on the chromatin signaling processes can aid in understanding the complex behavior of the whole signaling system. Further, by taking into account all available data during the reconstruction, these methods can discard potentially indirect interactions. This drives the research focus towards data-driven hypotheses on the global, molecular biological mechanisms. The aim of the present thesis is to apply these network reconstruction methods to understand the regulation of transcription at the chromatin level.

We address three main research objectives. First, we compare different network reconstruction methods theoretically and experimentally to reveal the advantages and disadvantages of each method when applied to genome-wide DNA-protein binding data. Second, we apply network reconstruction methods to derive testable hypotheses about the regulatory role of chromatin signaling at human promoters. To the best of our knowledge this is the first systematic computational study of chromatin signaling in human cells. Finally, we compare the chromatin signaling interactions across various transcriptional and regulatory states to provide further insight in the regulatory role of chromatin in distinct genomic and regulatory contexts.

1.2. Thesis outline

Following the introduction, we introduce in more detail the biological background in Chapter 2. We review the regulatory role of histone modifications and chromatin modifiers in transcription. The final paragraphs outline the genome-wide experiments available to study chromatin and transcription.

Chapter 3 presents the theoretical background on several network reconstruction methods. In Chapter 4, we motivate the use of the network reconstruction methods in the context of reconstructing chromatin signaling. We compare the different approaches on simulated data and genome-wide DNA-protein binding data in Chapter 5.

The first comprehensive set of ChIP-Seq data on chromatin modifier data in humans was published in 2011 [4]. We analyzed this set of proteins together with data on histone modifications. We provide hypotheses for mechanistic links between the histone modifications and the chromatin modifiers in humans. This project is described in Chapter 6.

The transcriptional and regulatory status of a genomic regions is related to different combinations of histone modifications. In Chapter 7, we investigate and compare the interactions of the chromatin modifiers at regions with different histone modification sets. We show that most interactions are context dependent. Further, individual members of known protein complexes also show state-dependent interactions. We conclude the thesis and discuss potential open questions and research directions in Chapter 8.

CHAPTER 2

Biological background

The cell has to make a variety of regulatory decisions at different stages of the transcriptional process. These include for example the choice of distinct regulatory or functional elements in the DNA, making the DNA accessible to regulatory proteins or assembling the necessary proteins for the initiation of transcription. The different regulatory decisions are carried out in a tightly time and space restricted manner. The chromatin environment plays an essential role in these processes and a large part of the regulatory effort is achieved through chromatin-associated proteins that transmit, memorize or interpret the appropriate regulatory signals. Together the complex interplay between these factors mediates the dynamic state of the genome that ultimately defines the cell's identity.

The following chapter serves the purpose of introducing the biological background relevant for understanding the scope and the remainder of the thesis. In Section 2.1, the proteins involved in packing the DNA into chromatin and interacting with chromatin are reviewed. Afterwards, Section 2.2 outlines the different levels at which the cell-specific regulation and use of the genome can be observed with a focus towards the role of chromatin. Finally in Section 2.3, the experimental procedures and their rationale in studying chromatin regulation on a genome-wide level are presented.

2.1. Chromatin components

As reviewed in [5], chromatin refers to the packing of the DNA along with proteins in the eukaryotic nucleus. The basic repeating unit, called *nucleosome*, is built by wrapping approximately 147 base pairs (bp) of naked DNA around an octameric protein complex consisting of *histone* proteins. Consecutive nucleosomes are linked by short stretches of linker DNA of varying length. The nucleosomes together with the linker DNA resemble a beads-on-a-string structure that is further coiled into the 30 nanometer (nm) fiber *in vivo*. Additional packing is achieved by looping of the 30nm fiber through scaffolding proteins. During mitosis these structures are even further compacted and stabilized into chromosomes. The different steps of packing are illustrated in Figure 2.1. The following section reviews different biological entities that are part of, interact with and/or regulate

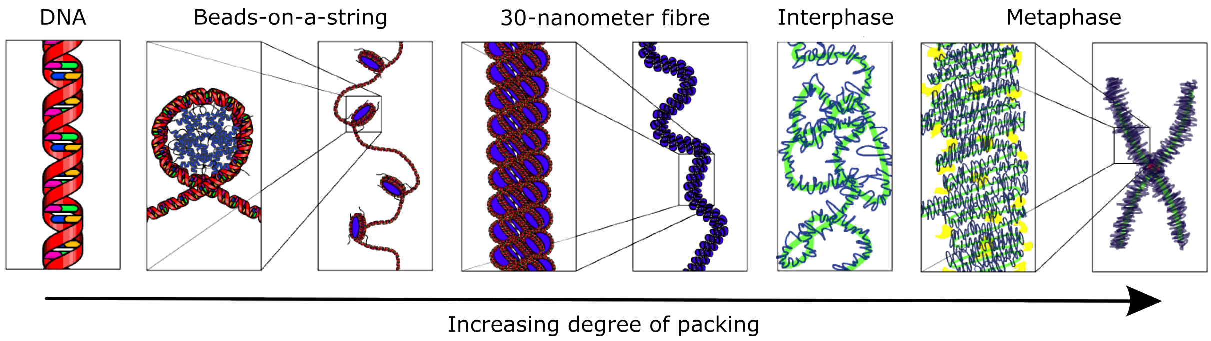

Figure 2.1.: *Illustration (adapted from [6]) of the different stages of DNA packing into chromatin. The DNA is wrapped around the histone core into nucleosomes. Consecutive nucleosomes are connected by the linker DNA forming the beads-on-a-string structure. Further compaction leads to the 30-nanometer fiber and finally to the characteristic metaphase chromosomes.*

chromatin. The main focus lies on histone proteins and chromatin modifiers.

2.1.1. Histones and histone modifications

The building blocks of chromatin are histones that together with DNA form the nucleosomes. Pairs of each of the four core histones H3, H4, H2A and H2B build an octameric protein complex. The interactions of the histone complex with the DNA lead to a wrapping of the DNA around the histone complex and thus to the formation of a nucleosome. Besides the four canonical histones, different histone variants exist that can replace the corresponding core histone. For example, histone H2A can be replaced by H2A.Z, H2A.X and macroH2A. The histone H3 can be replaced by e.g. H3.3. Histone variants have high similarity in their amino acid sequence to the corresponding core histone. However, they differ mostly in their structural features and their function from the canonical histone. The precise functional roles of these variants are still mostly unknown.

The unstructured aminoterminal parts of the histone proteins, called *histone tails*, stick out of the core nucleosome. They are thus accessible to enzymes that are able to modify specific residues within the histone tails. These post-translational, covalent modifications are called *histone modifications* (HMs) or *histone marks*. They include among others methylation (me), acetylation (ac), phosphorylation (p) and ubiquitinylation (ub) and affect for example lysines (K) and arginines (A). Typically, a particular histone modification is abbreviated by the histone name, the type and the position of the residue followed by the modification and the quantity of the modification added to the residue. For example, *H3K4me3* states that there are three methylgroups attached to the lysine at position four in the histone tail of histone H3.

The mere presence of a nucleosome affects the accessibility of a particular DNA region by the transcriptional machinery or by regulatory proteins such that nucleosomes have been initially thought to be general repressors of transcription. Acetylation of the histones loosens the contact to DNA and thus opens up the chromatin to allow transcription (reviewed in [7]). For methylation such an effect on the strength of the nucleosome binding is not known. However, the function of the different HMs seems to be more distinct. For example, different HMs recruit specific chromatin-modifying proteins, and thus are involved in specific signaling processes (see Section 2.2.3).

2.1.2. Cytosine modifications

The DNA itself can also be chemically modified by the addition of a methyl group to cytosines (5mC). In eukaryotes, DNA methyltransferases (DNMT) set and maintain the 5mC almost exclusively at CpG-dinucleotides. The TET proteins sequentially produce several other cytosine modifications. Due to the enzymatic activity of the TETs, 5mC can be oxidized into the 5-hydroxymethylation (5hmC) [8]. 5hmC can be catalyzed into 5-formylcytosine (5fC) which can then be turned into 5-carboxylcytosine (5caC) [9].
The different cytosine modifications have been implicated to function in various regulatory processes. The classical role attributed to 5mC is the repression of transcription and of repetitive elements (reviewed in [10]). 5hmC is present at repressed genes or genes that become active in later cell differentiation stages [11]. In contrast, 5fC was found at transcriptionally active genes [12]. Cytosine modifications can be recognized by distinct sets of proteins, including chromatin modifiers [13]. Hence, these modifications might function as signaling molecules analogous to histone modifications. However, the precise function of the cytosine modifications is still unknown.

2.1.3. Chromatin-modifying proteins

In the present thesis, chromatin modifiers refer to proteins harboring the capability to change the chromatin environment in proximity of their binding location. They can be distinguished by their enzymatic activity. Adenosine triphosphate (ATP) dependent chromatin remodelers are able to slide nucleosomes along the DNA and eject or exchange histones under ATP-hydrolysis. Methyl- and acetyl-transferases place methyl- or acetyl-groups, respectively, on specific amino acids on the histone tails. Demethylases and deacetylases remove these modifications. Other proteins (e.g. HP1 [14–16]) can bind to specific histone modifications and by, for example, complex formation compact chromatin. Multiple mechanisms exist to target the chromatin modifiers to specific sites in the genome. Inherent recruitment mechanisms are protein domains that recognize specific modifications on the histone tails (see table 2.1). The protein-protein interactions with sequence-specific DNA binding proteins, called transcription factors (see section 2.1.4), form another recruitment mechanism (e.g. [17, 18]). Other recruitment mechanisms include chromatin structure [19], non-coding RNAs (e.g. [20]) or DNA sequence composition (e.g. [21, 22]). Once a chromatin modifier is recruited to its target it can fulfill either its enzymatic activity or recruit other effector proteins, complex partners or the transcriptional machinery.

Table 2.1.: *Examples of typical protein domains that are able to recognize histone residues or modifications (reviewed in articles as indicated in column "Target").*

Domain	**Target**	**Examples**[1]
Bromo	lysine acetylation [23]	BRD4, ASH1L, KAT2A
Chromo	lysine methylation [24]	CBX2, CHD1, CBX3
PHD-type zinc finger	mainly lysine methylation [25]	PHF8, ASH1L, KDM2A

[1] Source: `http://www.uniprot.org/`

2.1.4. Other chromatin-interacting factors

Other factors that interact with chromatin include, for example, regulatory non-coding RNAs or transcription factors. Regulatory RNAs do not code for a protein but instead have regulatory functions. Especially long non-coding RNAs are relevant in the chromatin context as they for example recruit various proteins including chromatin modifiers [20]. Transcription factors (TF) are key regulatory proteins which usually function to promote or to reduce transcription depending on the context. Through a DNA binding domain, TFs can recognize particular small DNA sequences. TFs interact with chromatin in several ways. Many transcription factors bind mostly in nucleosome free regions [26]. However, pioneering factors (reviewed in [27]) bind to their DNA motif despite the presence of nucleosomes. These factors recruit chromatin remodelers which then open the chromatin for further regulation.

2.2. The regulatory role of chromatin

Chromatin has, apart from the mere packing of DNA in the nucleus, an essential function in the regulation of transcription. Thereby its role is dual: First, the strength of the packing directly influences the accessibility of the underlying DNA. This is important since various biological processes and proteins can fulfill their function only in a certain chromatin environment. Second, the histones, as part of the chromatin, can be read by other proteins and mark regulatory or functional regions of the genome. Together these aspects facilitate a cell-type specific, restricted use of the genome. The following paragraphs review the relation between the state of the chromatin and the different levels of transcriptional regulation.

2.2.1. Chromatin states

The local chromatin environment, called *chromatin state*, depends on the cell cycle stage and the transcriptional state of a particular genomic region. Disregulation of the chromatin state due to mutations in a particular chromatin modifier is associated with, for example, cancer [28]. Thus, it is critical for the cell to tightly regulate the particular chromatin state at each genomic location.

The chromatin can be broadly classified into two states [5]: *Euchromatin* and *heterochromatin*. Euchromatin refers to regions of loose chromatin, where the DNA is accessible to TFs and the transcriptional machinery, and is mostly associated with active transcription. In contrast, heterochromatin refers to tightly packed and associated with transcriptionally silent regions. In heterochromatin, much of the DNA is covered by dense nucleosome clusters preventing access to the transcription machinery.

The broad classification of chromatin states into euchromatin and heterochromatin can be further diversified by the presence of distinct sets of histone modifications [29, 30]. Each of these sets of histone modifications is associated with a particular chromatin structure and transcriptional or regulatory state. Further, these combinations distinguish different regulatory and functional elements in the genome.

The particular packing state is highly flexible, such that if the appropriate signals and proteins are present or absent a euchromatic region can turn into heterochromatin or *vice versa*. Chromatin-modifying enzymes or ATP-dependent chromatin remodelers are able

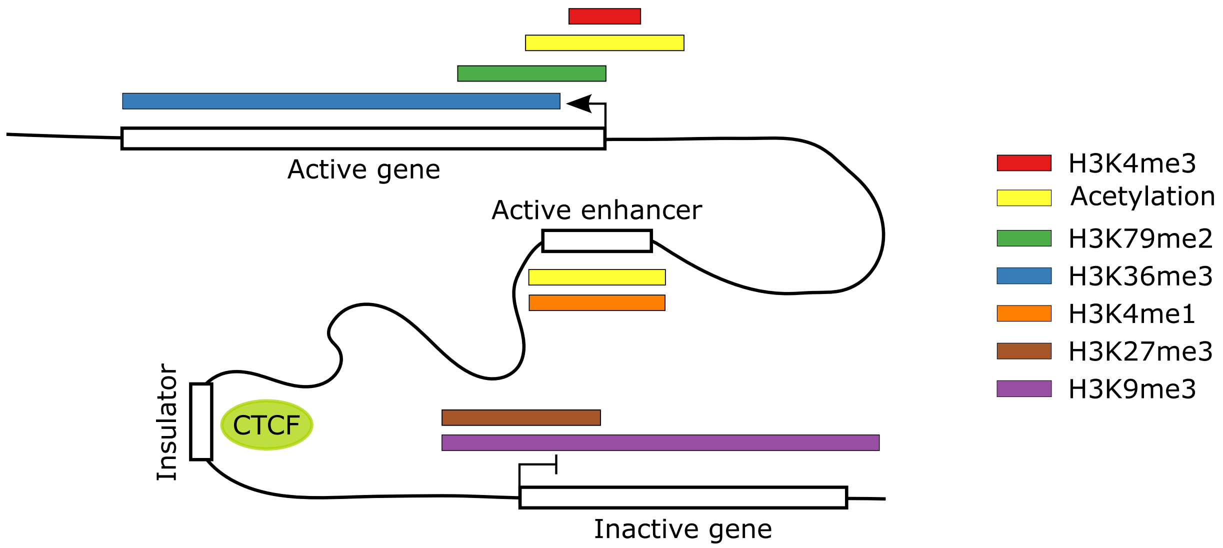

Figure 2.2.: *An illustration of the genomic elements across the genome and their relationship to distinct combinations of histone marks. Regulatory elements in the genome (sketched as a black line) are indicated as white boxes. Colored boxes indicate the distribution of a certain histone mark at the genomic region. The circle shows the protein CTCF.*

to switch the chromatin state of a particular region upon recruitment. These enzymes are recruited to defined regions of the genome allowing for a cell-specific regulation of the chromatin.

2.2.2. Regulatory elements

The combinatorial usage of regulatory elements is connected to the cell-type specific regulation of the transcriptional state of the associated gene [31, 32]. Examples of regulatory elements and their distribution along the genome are sketched in Figure 2.2. Promoters are regions where the transcriptional machinery is assembled. Promoter-proximal regions are in the vicinity of transcriptional start sites (TSS) and harbor various TF binding sites. Distant regulatory elements (enhancers, silencer and insulators) can be located up to several kilo base pairs (kb) upstream or downstream of a gene. Under the current model these elements function by building chromatin loops with other regulatory elements [33, 34]. Enhancers have an increasing and silencers have a decreasing effect on the transcription of the associated gene. Insulators or boundary elements separate different chromatin states and are also involved in long-range chromatin interactions [35].
Several genome-wide studies relate the different regulatory elements with unique chromatin states that even depend on the activity of the elements (reviewed in [36, 37]). Figure 2.2 summarizes the knowledge on the connections between well-studied histone marks and regulatory elements or their activity. Active Promoters are usually marked by H3K4me3 and histone acetylation at the TSS. H3K79me2 marks initiation in the first exon of the gene body. Actively transcribed regions are marked by H3K36me3. Poised or bivalent promoters, which are promoters whose final transcriptional state are undefined, are marked by H3K4me3 and H3K27me3. Repressed genes are usually marked only by H3K27me3 or in heterochromatic regions by H3K9me3. Active enhancers are charac-

terized by H3K4me1 and H3K27ac, while inactive enhancers might harbor H3K4me1 or H3K27me3. In summary, the different histone modifications clearly distinguish regulatory elements and thus might be seen as localized regulatory signals embedded in the chromatin.

2.2.3. Chromatin signaling

The discovery of different combinations of histone marks that define various chromatin states and proteins that can read off those combinations led to the *histone code hypothesis*. The histone code hypothesis states that "multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream functions" [38]. The histones were thus not only recognized as packing elements to build a particular chromatin structure but rather as signaling molecules. As such they recruit particular chromatin-modifying enzymes and would trigger a particular down-stream event. However, the statement that histone modifications act in combinatorial fashion to define a particular outcome currently has little evidence [39]. Instead, the different chromatin modifiers interact with the different histone modifications sequentially or in parallel and cross-talk with other cellular processes. The observations of combinatorial patterns might thereby be a side effect of obtaining stability of the signal. This observation led to a more dynamic view of the chromatin-related processes which now have to be embedded into the whole cellular signaling. The complex interactions involving the different chromatin components are called chromatin signaling interaction in the remainder of the thesis.

2.3. Experimental methods to study gene regulation

Various experimental techniques are available to study the different aspects of transcription and its regulation. A central part of these techniques is DNA sequencing which refers to reading out the particular chain of nucleic acids in a DNA fragment of interest. Each of these experimental techniques complements the sequencing step depending on the particular question with enrichment or conversion steps (illustrated in Figure 2.3). The following paragraphs describe the general principle of high-throughput sequencing techniques and the applications of DNA sequencing within the different protocols.

2.3.1. High-throughput DNA Sequencing

The basic principle behind all DNA sequencing techniques is the copying of a single-stranded DNA template into a double stranded fragment. Thereby the techniques exploit the principle of base complementarity such that in the newly synthesized strand a certain nucleotide is preferentially added depending on the template. Once a new nucleotide is added, a specific signal (e.g. a fluorescence light or the stop of the copying process) is read off by the sequencing machines allowing the reconstruction of the initial sequence. The DNA sequencing techniques have been rigorously improved for speed, cost and base coverage over the last decades. Thus, apart from their original application of sequencing the genome, DNA sequencing techniques are now also widely applied to study genome-wide transcriptional regulation.

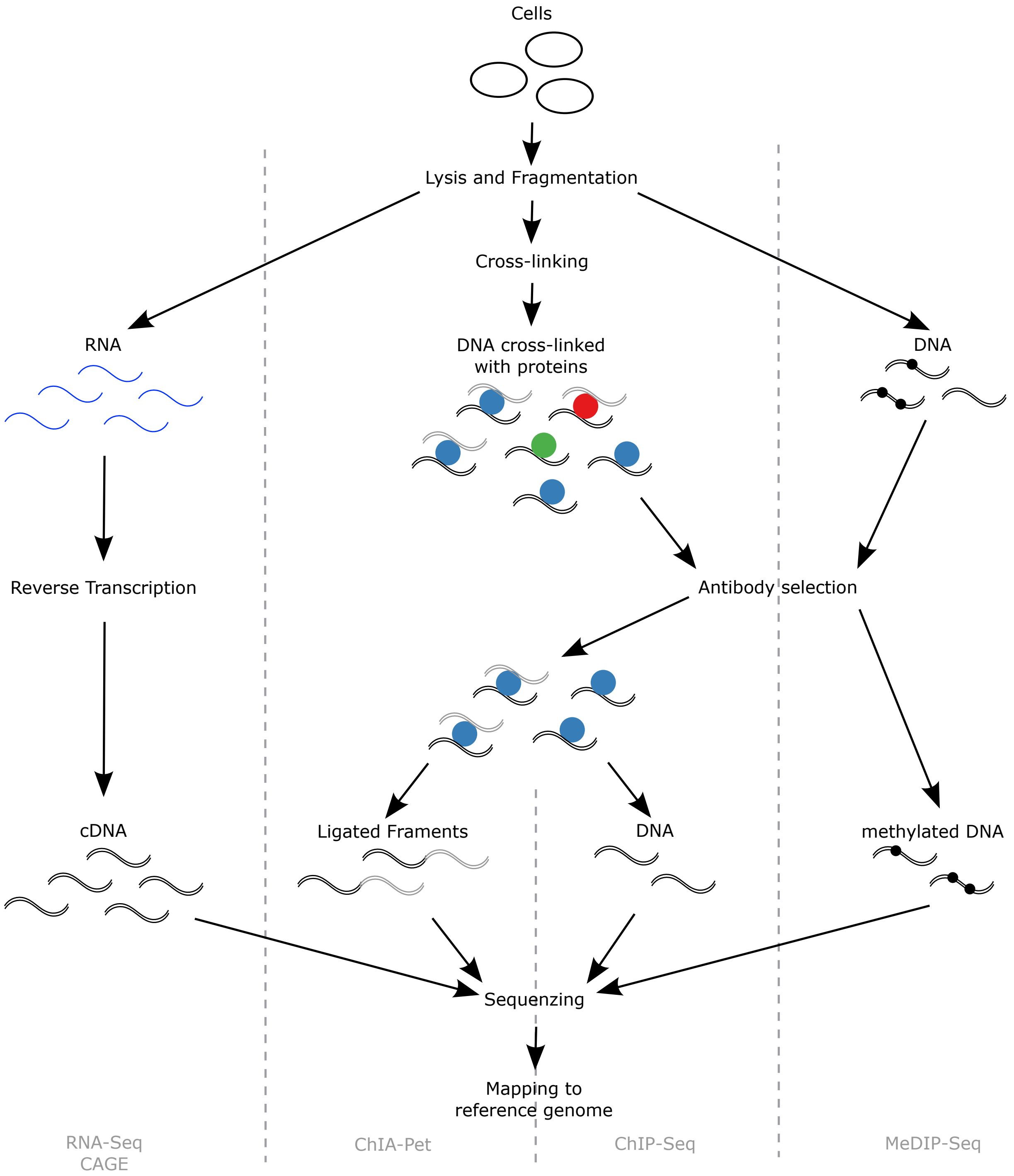

Figure 2.3.: *An illustration of the steps in the different experimental protocols to study gene regulation.*

2.3.2. RNA abundance

High-throughput sequencing-based protocols for measuring RNA abundance characterize and measure the steady state level of transcripts in the cells. The detected abundance levels thereby reflect transcription as well as degradation of the transcripts in a cell population. These sequencing-based techniques have almost completely replaced traditional Microarray studies which are by design limited, for example, in the detection of unknown transcripts and small regulatory RNAs or in the ability to distinguish particular transcript isoforms (reviewed in [40]).

The different protocols can be generalized to the following common steps: First, the sequencing library is prepared by enriching for the transcripts of interest. Next, the selected transcripts are reverse transcribed into cDNA which is then passed on to the sequencing machines. The resulting sequence reads can be mapped to a reference genome using computational tools.

The Cap Analysis Gene Expression (CAGE) technique is an example of such an high-throughput sequencing-based protocol [41, 42]. In the CAGE protocol only full length transcripts with a 5'-cap are selected for library preparation and the cDNA is produced only from short regions at the 5'end of these transcripts. From CAGE data one can thus, in contrast to other protocols (e.g. RNA-Seq), reliably identify the precise position of the transcriptional start site.

2.3.3. Protein-DNA binding events

Chromatin-immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) can be used to detect protein-DNA binding events [43–45]. The protocol starts with cross-linking the proteins to the DNA by exposing the cells to UV-light or formaldehyde. Afterwards, the DNA is fragmented and then incubated with antibodies against the proteins of interest. This marking is then used to enrich for DNA fragments that are bound to the proteins of interest. Finally, the proteins are washed off and the resulting DNA fragments are sequenced.

Another protocol to detect protein-DNA binding events is DamID [46]. Here, instead of a protein-specific antibody, an artificial fusion protein built from the protein of interest and an adenine specific methyltransferase is used. This protein introduces an artificial adenine methylation in the proximity of the DNA binding regions of the protein of interest. After sequencing the genome and comparing it to the reference genome, the regions of harboring artificial adenine methylation are identified.

2.3.4. Cytosine modifications

Several techniques (reviewed in [47]) are available to detect methylated cytosines in the genome. Probably, the most popular one is Bisulfite-Sequencing, where due to bisulfite treatment modified cytosines are converted into tymines and identified after sequencing by comparing to a reference genome. However, this technique does not distinguish between the classical 5mC-modification and 5hmC, 5caC or 5fC. Another technique based on methylated DNA immunoprecipitation (MeDIP) circumvents this problem. A antibody against the specific cytosine modification is used to enrich for fragments containing modified cytosines and to discriminate between the different cytosine modifications. The remainder of the protocol is very similar to ChIP-Seq. However, the usage of an antibody also leads to a disadvantage of MeDIP in comparison to Bisulfite-Sequencing. The resolution depends on the fragment size and as such is much lower than with Bisulfite-Sequencing.

2.3.5. DNA-DNA contacts

The genome-wide 3-dimensional organization of DNA can be measured by a plethora of different high-throughput methods (reviewed in [48]). In principle, DNA regions that are close in space are ligated to each other, making it possible to sequence pairs of interacting DNA regions. The difference in the various methods lies in the amount of pairs that can be analyzed at once. With a technique called Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-Pet) only interactions where a specific protein of interest is involved are sequenced [49]. The selection is again achieved by applying a specific antibody against the protein of interest after ligation. This technique reduces the noise resulting from non-functional or non-specific contacts by restricting the discovered interactions to regions where a protein is bound.

CHAPTER 3

Computational methods for reconstructing interaction networks

The task of reconstructing an interaction network, also called *reverse engineering* or *network inference*, aims at identifying the specific, statistically relevant associations between the components of a complex system from observed data. Network reconstruction methods have been applied to various problems, e.g. the reconstruction of protein-protein interaction [50], gene regulatory [51] or co-evolution networks [52]. Recently [53–55], different network reconstruction methods have been used to reconstruct chromatin-related signaling networks from DNA-protein binding data.

In the following chapter we review the task of reconstructing interaction networks. We introduce the notation used throughout this chapter in Section 3.1 and formulate the task of reconstructing interaction networks mathematically in Section 3.2. Section 3.3 summarizes the mathematical background of relevance networks. In Section 3.4 we introduce the concept of conditional independence and review different methods that build upon this concept and have been applied in the context of reconstructing chromatin-related interaction networks in Section 3.5-3.7.

3.1. Notation

Interaction networks are usually represented by a graph $G = (V, E)$ that summarizes the topology of the network. Here, V is a set of nodes and $E \subseteq V \times V$ is a set of edges connecting these nodes. Each node v_i in the set V represents a network component (e.g. a gene or a protein). An edge $(v_i, v_j) \in E$ denotes an interaction between the two nodes v_i and v_j (e.g. gene regulatory interactions or protein-protein interactions). These edges can be either directed or undirected, which means that $(v_i, v_j) \neq (v_j, v_i)$ or $(v_i, v_j) = (v_j, v_i)$, respectively. Depending on the type of edges, the interpretation of the edges changes (as discussed in section 4.4).

In the remainder of this thesis we will adopt a probabilistic framework where we associate a random variable X_i to each node v_i. The set of nodes can then be modeled as a multivariate random variable $X = (X_1, ..., X_N)$. Given some data D on the network, each measurement or observation of the nodes in the network in this framework constitutes a realization $(x_1, ..., x_N)$ of the random variable X.

3.2. Objective

Given the data D, the goal of reconstructing interaction networks is to find a model M that describes the relationships between the X_i's such that it best explains the observed data D. The network reconstruction process can be separated into structure learning and parameter estimation of model M. Structure learning in this context refers to selecting relevant edges between the nodes and is a combinatorial optimization problem: if there are N nodes there are $\binom{N}{2}$ possible edges. The various combinations of these edges lead to an exponential number of possible networks. The methods thus need to efficiently decide where an interaction is reasonable given the data. The parameter estimation learns the parameters of the interactions in the network model from the data. This is either done concurrently with the structure learning or done after a structure is fixed.

3.3. Relevance networks

One of the most intuitive ways of defining interactions between variables is to measure the pairwise association or similarity between two random variables. These pairwise association measures used in this context include, for example, the Pearson correlation coefficient [56] or the mutual information [57]. Given all empirical pairwise measurements of association, to construct the interaction network a threshold is applied to the association measures. The assumption behind this is that high-scoring interactions show relevant edges in the network.

3.3.1. Pairwise Pearson correlation

The intuition underlying the use of the pairwise Pearson correlation for network reconstruction is based on properties of the Pearson correlation coefficient. Under a multivariate Gaussian distribution we can assume that, if the Pearson-correlation coefficient

$$\rho_{X_i,X_j} = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)var(X_j)}} \tag{3.1}$$

between two Gaussian random variables X_i and X_j is zero, these two variables are *independent*. Thus, two nodes get connected in the network only if the empirical correlation coefficient between them is non-zero. Due to the properties of the Pearson-correlation coefficient this approach is most appropriate for detecting linear dependencies. The Spearman's correlation coefficient can identify monotonic relationships by computing the correlation coefficient on ranked data.

3.3.2. Pairwise mutual information

Similar to pairwise correlation, mutual information can detect pairwise associations in the case of discrete random variables. The mutual information between two discrete random variables X_i and X_j each having possible outcomes x and y, respectively, is defined as

$$MI(X_i, X_j) = \sum_x \sum_y P_{X_i,X_j}(x,y) log \left(\frac{P_{X_i,X_j}(x,y)}{P_{X_i}(x)P_{X_j}(y)} \right). \tag{3.2}$$

This measure obtains a value of zero if X_i and X_j are independent, i.e. only if the joint probability distribution $P(X_i, X_j)$ equals the product of the marginal distributions $P(X_i)$ and $P(X_j)$.

3.4. Conditional Independence

Using the pairwise association measures described in section 3.3 it is difficult to distinguish between different network structures underlying the observed data. For example, the situation where a variable X_1 interacts with X_2 that in turn interacts with X_3 is indistinguishable by simply measuring the pairwise association from the situation where X_1 interacts with both, X_2 and X_3. However, if data on X_1 is given, we can account for this information using the concept of *conditional independence*. In the following, we review conditional independence and its relation to the partial correlation coefficient based on [58].

Conditional independence for two random variables X_i and X_j given the set of remaining random variables $Z = X \setminus \{X_i, X_j\}$ holds, if and only if

$$P(X_i, X_j|Z) = P(X_i|Z)P(X_j|Z) \tag{3.3}$$

where $P(X_i|Z)$ refers to the conditional probability of X_i given Z. This means that, knowing Z, X_j does not contribute any new knowledge about X_i and *vice versa*. Intuitively this states that Z can explain the association between X_i and X_j.

Note, that if the random variable X follows a multivariate Gaussian distribution, conditional independence holds if and only if the *partial correlation coefficient* $\rho_{X_i,X_j|Z}$ equals zero. The partial correlation coefficient is thereby defined similarly to the Pearson correlation between X_i and X_j conditioned on Z

$$\rho_{X_i,X_j|Z} = \frac{cov(X_i, X_j|Z)}{\sqrt{var(X_i|Z)var(X_j|Z)}} \tag{3.4}$$

For an interpretation, we can think of the partial correlation coefficient as the Pearson correlation coefficient between the residuals of X_i and X_j after regressing each of them on Z (see also section 3.5.2).

3.5. Gaussian graphical models

In the multivariate Gaussian setting, the concept of conditional independence is used to construct the graph structure of undirected *Gaussian graphical models*. A Gaussian graphical model consists of a graph $G = (E, V)$ and a probability distribution P. The edges in E represent the relationships among a set of random variables $\{X_i, ..., X_n\}$ of which each is associated to a node in V. Together with the associated probability distribution P, the graph G defines the full joint distribution of the variables. Thereby, missing edges $(X_i, X_j) \notin E$ represent conditional independence between the two corresponding variables given all other variables in the graph (Markov property). These independence assumptions restrict the possible probability distributions associated to the graph. Hence, this representation not only gives a direct translation to interaction networks but also simplifies the computation of the parameters of the distribution and increases the robustness of the inference from data. The following section describes multiple ways to calculate the graph structure from observed data.

3.5.1. Covariance selection

Efficient ways of inferring the conditional independence restrictions from observed data exist for graphical models representing a multivariate Gaussian distribution. These approaches are called *covariance selection*. They build upon the definition of conditional independence in the multivariate Gaussian setting (see section 3.4). Computing the partial correlation coefficient (see equation 3.4) can be done efficiently by inverting the covariance matrix as sketched in the following. The details of the proof can be found in [58, 59].
Assume a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$ with invertible covariance matrix Σ that can be partitioned as follows

$$\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \tag{3.5}$$

where A and B partition the set of variables X into distinct subsets X_A and X_B, i.e. $A \cup B = \{1...n\}$ and $A \cap B = \emptyset$. For example, let $A = \{X_i, X_j\}$ and $B = Z = X \setminus \{X_i, X_j\}$ as above. The conditional distribution of X_A given $X_B = x_B$ follows a Gaussian distribution with parameters as follows:

$$\mathcal{N}\left(\Sigma_{AB}\Sigma_{BB}^{-1}x_B, \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right) \tag{3.6}$$

From this we can observe that A and B are independent if and only if $\Sigma_{AB} = 0$, since it then holds that the mean $\mu_{A|B}$ and the covariance matrix $\Sigma_{A|B}$ of the conditional distribution are not influenced by the variables in B. Further, defining the concentration matrix as $K = \Sigma^{-1}$ and partitioning it as follows

$$K = \begin{bmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{bmatrix} \tag{3.7}$$

we find that

$$\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} = K_{AA}^{-1}. \tag{3.8}$$

This identifies K_{AA} as the concentration matrix of the conditional distribution of X_A given $X_B = x_B$. By observing similarly to the derivation of equation 3.8 that $-\Sigma_{AB}\Sigma_{BB}^{-1} = K_{AA}^{-1}K_{AB}$, it follows that A and B are independent if only if $K_{AB} = 0$
Since the concentration matrix of the conditional distribution is

$$K_{AA} = \begin{bmatrix} k_{ii} & k_{ij} \\ k_{ji} & k_{jj} \end{bmatrix} \tag{3.9}$$

and it holds that

$$\Sigma_{A|B} = K_{AA}^{-1} = \frac{1}{det(K_{AA})} \begin{bmatrix} k_{jj} & -k_{ij} \\ -k_{ji} & k_{ii} \end{bmatrix} \tag{3.10}$$

the conditional covariance $cov(X_i, X_j|X_B) = \frac{-k_{ij}}{det(K_{AA})} = 0$ if and only if $k_{ij} = 0$. Thus, a direct connection of the concentration matrix K to the partial correlation coefficients arises when scaling K to have all diagonal elements equal to one. Then it follows for all off-diagonal elements that

$$-\frac{k_{ij}}{\sqrt{k_{jj}k_{ii}}} = \frac{cov(X_i, X_j|X_B)}{\sqrt{var(X_i|X_B)var(X_j|X_B)}} = \rho_{ij|B} \tag{3.11}$$

giving the partial correlation coefficient $\rho_{ij|B}$ for two variables X_i and X_j. Hence, equation 3.11 tells us that these scaled entries of the concentration matrix, i.e. the inverse of the covariance matrix, are the partial correlation coefficient between two variables conditioned on all the remaining variables.

In summary, equation 3.8 states that the concentration matrix of the conditional distribution can be obtained by simply deleting the rows and columns corresponding to variables in B from the full concentration matrix of the joint distribution of X. And further, that we can define Gaussian graphical models by appropriately restricting elements in the concentration matrix to zero. To infer the conditional independence restrictions from observed data, least-square or maximum-likelihood approaches are employed. Further, for computational reasons and because it is convenient for the interpretation most recent methods assume sparsity and directly incorporate this assumption in the model fitting (e.g. [60, 61]).

3.5.2. Neighborhood selection

Meinhausen and Bühlmann [60] put forward an approximation approach called *neighborhood selection*. Here, the covariance selection problem is reduced to the problem of finding the smallest subset of nodes X_A, called *neighborhood*, such that a node X_i is conditionally independent from the remaining variables given the neighborhood. Defining X_B as $X \setminus \{X_i\}$ we get from equation 3.6 that the conditional expectation of X_i given $X_B = x_B$ is a univariate Gaussian distribution with:

$$\begin{aligned} E(X_i|X_B = x_B) &= \Sigma_{iB}\Sigma_{BB}^{-1}x_B & (3.12)\\ &= \textstyle\sum_{j\in B}\beta_{ij}x_j & (3.13)\end{aligned}$$

This resembles a regression of X_i on all remaining variables with partial regression coefficients $\beta_{ij} = -K_{ij}/K_{ii}$. It follows that the zero partial regression coefficients correspond to all $K_{ij} = 0$. The neighborhood X_A of node X_i is then given by all X_j with $\beta_{ij} \neq 0$, i.e. $X_A = \{X_j|\beta_{ij} \neq 0, j \neq i\}$.

Thus, the task of covariance selection is divided into finding the neighborhood for each node individually given the data. The coefficients in the fitted regression model are then interpreted as the rows in the concentration matrix. In [60] the neighborhood is estimated from the data by performing a Lasso regression for which the objective function is formulated as follows

$$\hat{\beta}_i = argmin_\beta||x_i - \beta x_A||_2^2 + \lambda||\beta||_1 \tag{3.14}$$

where $||\cdot||_1$ is the L_1-norm, also called Lasso penalty, and $||\cdot||_2$ is the L_2-norm. The regularization parameter λ penalizes the deviation of the coefficients from zero. As a consequence, large λ shrink non-informative coefficients towards zero.

The estimates of the entries in the covariance matrix do not necessarily result in a symmetric matrix. To make the matrix symmetric, the authors in [60] define the entry K_{ij} to be zero if either X_i is not in the neighborhood of X_j or vice versa, or alternatively only if both is true.

3.5.3. Graphical lasso

Several other approaches, including Graphical lasso [61], try to estimate the exact solution to the maximum likelihood fit of the covariance matrix. The key observation is that the covariance sub-matrix $\hat{\Sigma}_{BB}$ estimated by the neighborhood selection approach needs to be equal to the observed covariance sub-matrix to result in the maximum likelihood estimator by the neighborhood selection, which is not the case in general (as reviewed in [61]). The Graphical lasso now approaches this problem by iteratively solving the Lasso problem for each variable and updating the estimated covariance matrix until convergence. Thereby, at each step the procedure takes into account the current estimate of the covariance sub-matrix $\hat{\Sigma}_{BB}$. This leads to sharing information between the different regression problems. Together with a fast gradient descent method for solving the individual Lasso problems the outlined procedure results in the maximum likelihood estimator in moderate computation time.

3.5.4. Sparse partial correlation networks

Sparse partial correlation networks [54] build upon covariance selection but modify it in several ways. The overall goal of these modifications is to make the selection of the edges as precise and robust as possible. Here we only describe these modifications and refer to the original publication [54] for details.

The first modification is the application of the rank transformation to the data. This is intended to make the method less sensitive to the distribution of the data, which is in many applications non-Gaussian. Ranked data is uniformly distributed in the interval $[0, N]$, where N is the number of samples, and, as argued in [54], can be approximated by a flat Gaussian distribution making the data suitable for covariance selection. Additionally, by ranking the data robustness to technical variation and outliers is increased. However, the quantitative information in the data is partially lost and the results have to be interpreted differently, because the method now detects associations between ranks.

Further, in the sparse partial correlation network method a cross-validation scheme, in which only edges with strong support in the data are called, enforces sparseness on the recovered network. This is done by masking entries in the partial correlation matrix based on the following approach. In each cross-fold, a partial correlation matrix P is constructed on the training data. For each variable X_i, all variables having a non-zero partial correlation with X_i are selected from the matrix P and used for predicting X_i. Based on the prediction error averaged over all variables a global cutoff on the partial correlation coefficients is determined and all interactions above this cutoff will be set to zero. Finally, all interactions that are zero in a certain number of cross-folds are masked in the final partial correlation matrix.

3.6. Bayesian networks

In the following section, we review the use of Bayesian networks to infer interactions between variables from observed data (based on [62]). *Bayesian networks* encode the relationships between the variables as a directed, acyclic graph $G = (V, E)$. As before (see section 3.5), the graph illustrates the conditional independence structure between the variables and defines the full joint probability distribution. The parents $Pa(X_i)$ of a variable X_i are those variables with $(v_j, v_i) \in E$ that make X_i conditionally independent of all remaining variables that are non-descendants of X_i, i.e. all X_k with $(v_i, v_k) \notin E$. Thus, it is sufficient to look at the conditional distributions of X_i on its parents $P(X_i)$, called *local probability distribution*, to obtain the full joint probability distribution. Using this property and the probability chain rule, the joint probability can be computed by factorizing it according to the graph structure as follows:

$$P(X_1, ...X_N) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \tag{3.15}$$

From equation 3.15 we can observe that Bayesian networks allow the implementation of conditional independencies of all orders. This is in contrast to the Gaussian graphical models described in section 3.5 that assume full conditional independence. And further, we can observe that the joint distribution is defined only by the local probability distributions that in turn are defined only through a parent-child specific parameter set θ_i. For reasons of computational efficiency one usually assumes a Gaussian distribution in the case of continuous data or a multinomial distribution in the case of discrete data for the local probability distributions. The following section describes how to efficiently infer the parent-child relationships from observed data using a Bayesian approach.

3.6.1. Bayesian Score

To infer the structure of the Bayesian network given some data D several candidate graph structures are usually compared based on a score that indicates how well a graph models the observed data. The Bayesian score, for example, evaluates the posterior probability $P(G|D)$ of a candidate graph G. This probability can be approximated using Bayes' theorem

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \tag{3.16}$$

where $P(D|G)$ is the likelihood of the data given a proposed graph and $P(G)$ is the prior probability of the graph. The probability $P(D)$ is usually omitted since it does not influence the comparison of different graphs. If prior information on the graph structure is available it is possible to incorporate that prior knowledge through $P(G)$. Using a uniform prior $P(G)$ the score only depends on the likelihood $P(D|G)$.

The likelihood $P(D|G)$ is defined by the specific parameter sets θ_i for each local probability distribution i. Thus the maximal score that can be achieved for a given graph G is found by optimizing all parameter sets θ. In the Bayesian setting one does not compute the score on a single set of parameters, but instead marginalizes the likelihood over all parameter sets:

$$P(D|G) = \int_{\theta} P(D|\theta, G)P(\theta|G)d\theta \tag{3.17}$$

As a result the influence of a particular choice of θ on the posterior probability vanishes. This step is intended to prevent over-fitting, which would otherwise cause complex models to have a better score than simpler ones only due to the larger number of parameters available for modeling the data. Under certain circumstance (as reviewed in [62]), it is possible to compute the marginal likelihood analytically. The marginal likelihood can then be decomposed into variable-wise contributions that depend on the corresponding local probability distributions.

3.6.2. Graph structure selection

Since Bayesian networks allow to model any order of conditional independence it is computationally infeasible to evaluate all possible graph structures. To propose candidate graphs most learning algorithms make use of heuristic search algorithms, e.g. simulated annealing or greedy hill-climbing, that make elaborate choices of iteratively chosen candidate graphs. Thereby in each iteration a slightly modified version from the current graph is proposed and accepted under certain conditions. These candidate structures are scored against each other using, for example, the criterion in equation 3.16 and after the stopping criterion is fulfilled the highest scoring graph is returned. Finally, since it is often the case that several high scoring graphs exists, sampling techniques, e.g. Bootstrapping [53], are employed to retrieve confidence scores for the edges.

3.7. Maximum-entropy approach

Another approach for reconstructing interaction networks is based on the principle of maximum entropy, which we will explain in the following chapter based on [55, 63]. In contrast to the sections 3.5 and 3.6, we assume a discrete multivariate random variable $X = (X_1, ..., X_N)$. But again we model the full joint probability distribution $P(X_1, ..., X_N)$ to then infer the interactions among the variables.

In the *maximum entropy approach* the objective is to learn the probabilities p_m for each possible observation x_m with $m \in \{1, ..., M\}$ from some summary of the observed data. In the following chapter, we will see that this task can be formulated as an optimization problem and can be used to reconstruct interaction networks from observed data.

3.7.1. Entropy

Entropy measures the uncertainty represented in a probability distribution. Assuming a discrete multivariate random variable $X = (X_1, ..., X_N)$ with possible events x_m and associated probabilities p_m $(m \in \{1, ..., M\})$, the Entropy H of the probability distribution according to Shannon is given by

$$H = -\sum_{m=1}^{M} p_m log(p_m) \tag{3.18}$$

3.7.2. Entropy maximization

We can use the Entropy measure to estimate the probabilities p_m under the constraint that all p_m sum up to 1 by formulating the following optimization problem:

$$\begin{aligned} \max_{p_m} \quad & -\sum_{m=1}^{M} p_m log(p_m) \\ \text{s.t.} \quad & \sum_{m=1}^{M} p_m = 1 \end{aligned} \tag{3.19}$$

The solution to this constrained optimization problem is given when all p_m are equal. Hence, the Entropy measure is such that, when no additional information is given, maximizing the entropy with respect to the probability distribution leads to the uniform distribution.

The estimate of the probability distribution in equation 3.19 can be extended to require agreement of the distribution with the observed data. In more detail, assume we are given some values F_k with $k \in \{1, ..., K\}$ that are the observed outcome of the following function:

$$F_k = \sum_{m=1}^{M} p_m f_k(x_m) \tag{3.20}$$

where f_k is some function of the possible observations x_m. The equations in 3.20 can be used directly as constraints in the optimization problem. Using Lagrange multipliers λ_k, it can be shown that the optimization problem in 3.19 under the constraints in 3.20 results in the following probability distributions:

$$p_m = \frac{1}{Z(\lambda_1, ..., \lambda_K)} exp(\sum_{k=1}^{K} \sum_{m=1}^{M} \lambda_k f_k(x_m)) \tag{3.21}$$

where $Z(\lambda_1, ..., \lambda_K)$ is a normalization factor, called *partition function*, that ensures the proper normalization of the probability distribution. The parameters λ_k depend on the constraints in equation 3.20 and can be computed by plugging equation 3.21 into the constraints 3.20. For simple applications the λ_k are easy to compute, but with more constraints this task is not feasible analytically.

3.7.3. Network inference

We can use the maximum entropy solution given in equation 3.21 to infer interactions between the variables X_i by choosing appropriate constraints. For example, we can require that the expected value of a random variable X_i in our model should agree with the observed expected value μ_i. This is achieved by formulating the following constraints for each random variable X_i:

$$\mu_i = \sum_{m=1}^{M} p_m x_{mi} \tag{3.22}$$

where x_{mi} indicates the outcome for X_i in the event x_m. Similarly, we can model associations between two variables X_i and X_j with the following constraints for each pair of random variables X_i and X_j:

$$\mu_{ij} = \sum_{m=1}^{M} p_m x_{mi} x_{mj} \tag{3.23}$$

Using the same approach as in section 3.7.2, these constraints then lead to the following maximum entropy solution:

$$p_m = \frac{1}{Z} exp(\sum_{i=1}^{N} \sum_{m=1}^{M} \lambda_i x_{mi} + \sum_{i=1}^{N} \sum_{j=i}^{N} \sum_{m=1}^{M} \lambda_{ij} x_{mi} x_{mj})) \tag{3.24}$$

where Z is the partition function depending on the λ_i and λ_{ij}. The solution to the maximum entropy problem leads to an interpretation of the λ_i as the preference for a particular outcome of X_i. Similarly, the λ_{ij} can be interpreted as the strength of the interaction between the two variables X_i and X_j.

In principle one could also take into account higher order interactions between the random variables. But the more constraints there are the more parameters have to be estimated. Already in the case of modeling pairwise interactions in many applications heuristics have to be applied [55, 63]. These heuristics are used to solve the optimization problem and compute the parameters λ_i and λ_{ij}.

CHAPTER 4

Network reconstruction for chromatin signaling

In Section 2 we have summarized the role of histone modifications and chromatin modifiers in transcriptional regulation. HMs and CMs seem to be part of a chromatin signaling network, but the precise interactions are usually unknown. Recently more and more genome-wide DNA-protein binding data of HMs and CMs was produced [4, 26] allowing the systematic inference of potential signaling interactions. The huge amount of data generated by these experiments calls for computational and visualization techniques that allow to represent the resulting interactions in a comprehensive way.

In the following chapter we motivate the use of the network reconstruction methods described in Section 3 to reconstruct chromatin signaling interactions. We give the general assumption underlying the detection of interactions from genome-wide DNA-protein binding data in Section 4.1. In Section 4.2 we summarize how the experimental data is pre-processed towards a data matrix suitable for the computational methods. We motivate the use of network reconstruction methods in the context of chromatin-signaling in Section 4.3. Finally, in Section 4.4 we discuss the interpretation of the resulting networks.

4.1. Central assumption

Genome-wide DNA-protein binding data (see Section 2.3.3) refers to a collection of sequences of reads, i.e. small segments of DNA sequences, from genomic locations that were bound by the protein of interest. These reads are mapped computationally to a reference genome and thus the exact binding location of a protein can be determined. The more reads are detected at a specific location, the more cells had a protein bound in this region. The genome-wide distribution of reads gives a binding profile that is very specific to the protein of interest (see Figure 4.1.

Under the assumption that two proteins act in the same pathway or even interact physically we would expect to see similar binding profiles. Thus, a central assumption in reconstructing chromatin-related signaling from genome-wide DNA-protein binding data

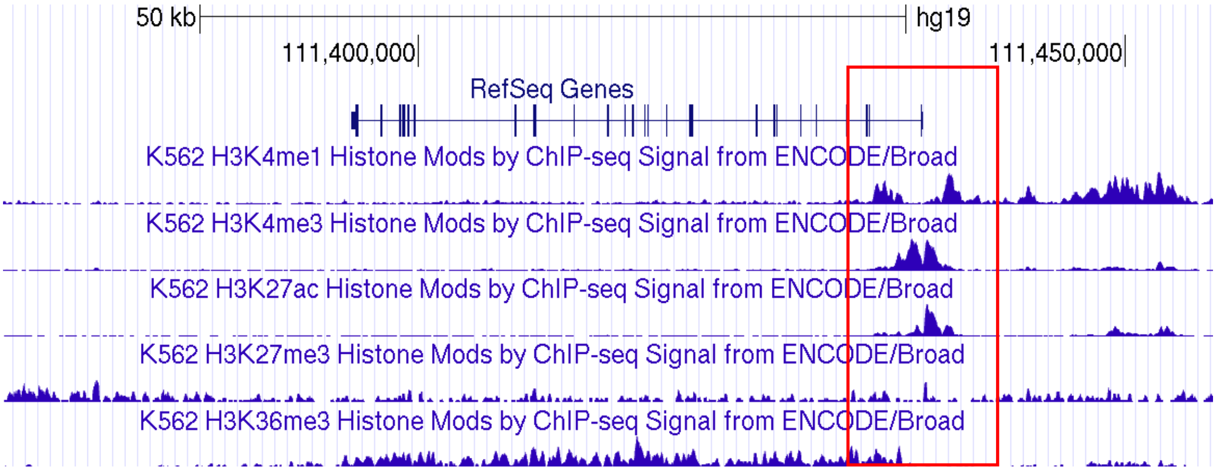
50 kb
hg19
111,400,000
111,450,000
RefSeq Genes
K562 H3K4me1 Histone Mods by ChIP-seq Signal from ENCODE/Broad
K562 H3K4me3 Histone Mods by ChIP-seq Signal from ENCODE/Broad
K562 H3K27ac Histone Mods by ChIP-seq Signal from ENCODE/Broad
K562 H3K27me3 Histone Mods by ChIP-seq Signal from ENCODE/Broad
K562 H3K36me3 Histone Mods by ChIP-seq Signal from ENCODE/Broad

Figure 4.1.: *Genome browser screenshot of the binding profiles of several ChIP-Seq experiments in a 50kb-long genomic segment. Top row gives the position in the hg19 genome. The next row shows an annotated gene. The following rows summarize the reads mapped to a particular position for each experiment. The red square highlights co-localization of several histone modifications.*

is that co-location suggests functional similarity. To infer interactions we thus need to measure the similarity between the binding profiles of two protein.

4.2. Data representation

An intuitive way of summarizing genome-wide DNA-protein binding data is to build a data matrix D where each row corresponds to a genomic region and each column corresponds to an experiment for a particular protein. All the different network reconstruction methods described in Section 3 build upon such a matrix. In this framework, each row in the data matrix D gives a measurement of the multivariate random variable $X = (X_1, ..., X_n)$ where each X_i represents a particular protein.

There are different ways to build the data matrix D that also depend on which method is going to be used. First, one could count the number of reads that fall into a particular set of regions, e.g. promoter regions. Second, these counts could be binarized using a cutoff to obtain a binary data matrix. Alternatively, one could for example identify peaks within the genomic regions (e.g. [53]).

There are advantages and disadvantages to each choice of the pre-processing method. Discretization is thought to make the assignments more stable and simplifies the computational steps in, for example, the Bayesian network and the Maximum-Entropy approach (see Section 3). However, the quantitative information is lost through this procedure. Further, the data is usually discretized into two states that are interpreted as protein-bound and protein-unbound. This does not necessarily reflect what is measured by, for example, ChIP-Seq data as these experiments measure the presence of a protein over a whole cell population. The count data preserves this quantitative information and can be used directly for inference with graphical models without loss of computational speed. However, these methods usually assume that the data follows a Gaussian distribution, which might not be fulfilled by the data. To circumvent this problem, it was suggested in [54] to use the rank of the binding scores for inferring the network because ranked data resembles a flat Gaussian distribution.

4.3. From co-localization to interaction networks

Given the data matrix D summarizing the binding profiles of HMs and CMs, we would like to infer the interactions between the HMs and the CMs. These interaction are usually represented as a graph (see Section 3.1). In the application of chromatin-signaling, each node in the graph represents an HM or a CM and each edge indicates a statistically significant co-location.

The interaction networks can be inferred by computing a score, which indicates the relationships between two variables, for each possible edge in the graph. The network reconstruction methods described in Section 3 all automatically deliver such a score from their model parameters. Depending on the network reconstruction methods used this edge score is quantitative or qualitative. For example, in the case of the pairwise association-based approaches the edge score is simply the correlation or the mutual information in the observed measurements D between any pair of variables.

4.4. Interpretation of the network

Some care has to be taken when interpreting the interaction networks derived from genome-wide binding data. The nature of the experiments limits the biological knowledge that can be derived from the detected interactions. We explain potential pitfalls in the following section.

Since the genome-wide DNA-protein binding data describes the presence of a protein at a given point in time, it has no direct information about the causes or the consequences of a binding event. The information captured by these experiments in a specific cell type thus only allows for the inference of co-locations. The resulting graphs is undirected. In contrast, directed graphs assign a direction to the edges that is usually used to reflect causal relationships. For the inference of causal relationships we would need to measure the dynamic properties of the system by, for example, perturbation experiments. However, these experiments are usually not available for HMs and CMs.

In the remainder of the thesis, with an interaction in the context of chromatin-signaling we refer to a statistically significant co-location of two proteins detected in the DNA-protein binding data. However, as indicated above the detected interactions do not necessarily imply physical contact, but result in data-driven starting points for deriving hypotheses about biologically relevant interaction partners. Additionally, the resulting networks can be used to get an overview of the interactions within a complex system of interest. From this overview it is possible to identify potential modules and pathways that can be used to understand the complex behavior of the system.

CHAPTER 5

Experimental comparison of network reconstruction methods

Each of the network reconstruction methods described in Section 3 makes different assumptions on the data and has a slightly different objective. In the following chapter, we are going to compare the different methods under various aspects that occur when working with ChIP-Seq data, e.g. noisy measurements. In Section 5.1, we measure the performance of the methods on simulated data. In this way, we obtain a controlled setting for which we know the true interactions and that the data on the nodes in the networks is complete. Next, we test the methods on a large publicly available data set from *Drosophila* that has been generated and processed by a single lab in Section 5.2. This offers the opportunity to deal with real biological data and study the effect of technical and biological noise on the methods. We conclude the chapter with a summary and discussion in Section 5.3.

5.1. A method comparison on simulated data

Properly evaluating the accuracy of the different methods is only possible if we know the true underlying network of the observed data. However, to the best of our knowledge there is no gold standard chromatin-signaling network available where the complete set of interactions is known. A possible data source for validation are protein-protein interaction (PPI) databases. However, for reasons explained in more detail in Section 5.2.5, PPIs are not a good gold-standard set, mostly because PPIs are incomplete and noisy themselves.

Additionally, to extract a network containing only true direct interactions it is necessary to observe all the proteins that are involved in the networks. In more detail, if the protein A that mediates an interaction between two others (B and C) is not observed, we are not able to infer that the interaction between B and C is indirect. Thus, the result largely depends on the completeness of the data set. Most available experimental data sets of CMs and HMs are not complete, rendering the distinction between indirect and direct interaction difficult. To remedy this situation, we study and compare the different methods on simulated data in the following section.

5.1.1. Simulated networks

As explained in Section 3, the available network reconstruction methods either assume a multivariate Gaussian distribution or use binarized data. Thus, we simulate multivariate Gaussian networks, which can be used directly by the methods that assume continuous data. For the remaining methods, we then binarize each sample in the simulated data into two groups by applying a cutoff defined as the average signal per sample.

The simulated networks are generated using the *qpgraph*-package [64] in R. In short, we define a random graph structure between n nodes with maximum number of edges s_{max}. The structure is created under the *Erdös-Rényi* graph model, in which the s_{max} edges are uniformly distributed between the nodes. From this, we define the covariance matrix such that an entry in the inverse of the covariance matrix is 0 if the corresponding pair of nodes is not connected by an edge in the simulated graph. Finally, we create m random observations from the simulated covariance matrix, which are normalized to have mean 0 and standard deviation 1. The standardization step serves to make the predicted weights in the different methods comparable.

5.1.2. Methods

We compare the different methods discussed in Section 3 by using the software or R packages as indicated in Table 5.1. As baseline methods we use simple empirical pair-wise Pearson correlation (Cor) for continuous and Mutual Information (MI) for binary data. The implementation of the other approaches are described in detail in the following.

The Partial Correlation (PC) and the Sparse Partial Correlation Networks (SPCN) are computed as described in Section 3.5. We use the same cutoff for masking entries in the SPCN as described in the original paper [54], i.e. in each cross-validation (CV) fold starting from the smallest partial correlation entry the entries are set to zero until a maximum decrease of 10% in the total prediction accuracy is observed. For retrieving the SPCN, an interaction in the partial correlation matrix is masked if 7 out of 10 CV-folds resulted in a zero entry.

Similar to the neighborhood selection method described in Section 3.5, we apply Elastic Nets (ENet) to reconstruct networks. Elastic Nets extends the Lasso regularization term on the regression coefficients by the L2-norm on the coefficients. As a consequence, similarly good predictors achieve similar weights [65]. In the context of chromatin signaling, where we assume that protein complex formation is reflected in the data, this seems more reasonable than choosing one complex member over the other.

Further, we apply the Graphical lasso (GL) approach. For both, Graphical lasso and Elastic Nets, we choose the penalty parameter for the regularization in a 10-fold CV. We choose the most stringent regularization parameter for which the mean prediction error is within one standard deviation of the parameter with minimal prediction error.

Similarly to the observations in [54] on partial correlations, we expect Elastic Nets and Graphical lasso to obtain mostly non-zero scores for potential edges in the case of a large number of observations. For this reason we also implement a sparse version of Elastic Nets (SENet) and Graphical Lasso (SGL). For SENet, we retain only those regression coefficients that exceed a threshold of one standard deviation from the mean of all regression coefficients in the prediction of each node. The rest of the scores is set to zero. Thus, the threshold can vary for each variable which is different from SPCN, where a global threshold is chosen. For SGL, we apply the same CV-based masking technique as in SPCN.

Finally, as an example for a Bayesian method, we apply Banjo as described in [53] to the binarized data. We draw 500 bootstrapping samples on which we run Banjo and count how often a certain edge (independent of the inferred direction) is recovered. The interaction score is then given by the fraction of bootstrapping samples in which the edge was detected.

The Maximum-Entropy approach (ME) is also applied to binary data as described in the original publication [55]. The regularization parameters are fitted as in [55] and we allow for pair-wise but no higher order interactions in the model. We additionally perform a 10-fold CV and the final score is given as the average score over the 10 CV-folds.

Table 5.1.: *Overview of the implementations of the network reconstruction methods studied in the method comparison. The first and second column summarize the methods and the abbreviations used throughout the chapter. The final column gives the main software or R package used throughout the thesis.*

Method	Abbreviation	Software/R package
Pearson correlation	Cor	R: stats
Mutual Information	MI	R
(Sparse) Partial Correlation	PC/SPCN	R: SPCNs[1]
(Sparse) Elastic Nets	ENet/SENet	R: glmnet
(Sparse) Graphical Lasso	GL/SGL	R: glasso
Bayesian network	Banjo	Banjo[2]
Maximum-Entropy	ME	matlab code[3]

5.1.3. Comparison on simulated data

We first discuss the general behavior and the different objectives of the methods on a single exemplary network with simulated data. For visualization purpose, we simulate a network with $n = 20$ nodes, $s_{max} = 40$ edges and $m = 5.000$ observations. The true network and the graphs resulting from the different methods are shown in Figure 5.1, where the scores of the interactions are indicated by the color intensity. Overall, all approaches show good results in recovering the true interactions. However, also some small differences become apparent as explained in the following.

Figure 5.1 indicates that, as expected, the Cor and MI approach recover many false positive interactions. Even though these interactions are false positives, the correlation coefficient or the mutual information are high rendering the identification of the true edges difficult. Mutual information, however, results in a wider distribution of scores compared to Cor, which results mostly in scores larger than 0.2. This might indicate a slightly

[1] `http://spcn.molgen.mpg.de/`

[2] `http://www.cs.duke.edu/~amink/software/banjo/`

[3] Kindly provided by Jian Zhou [55]

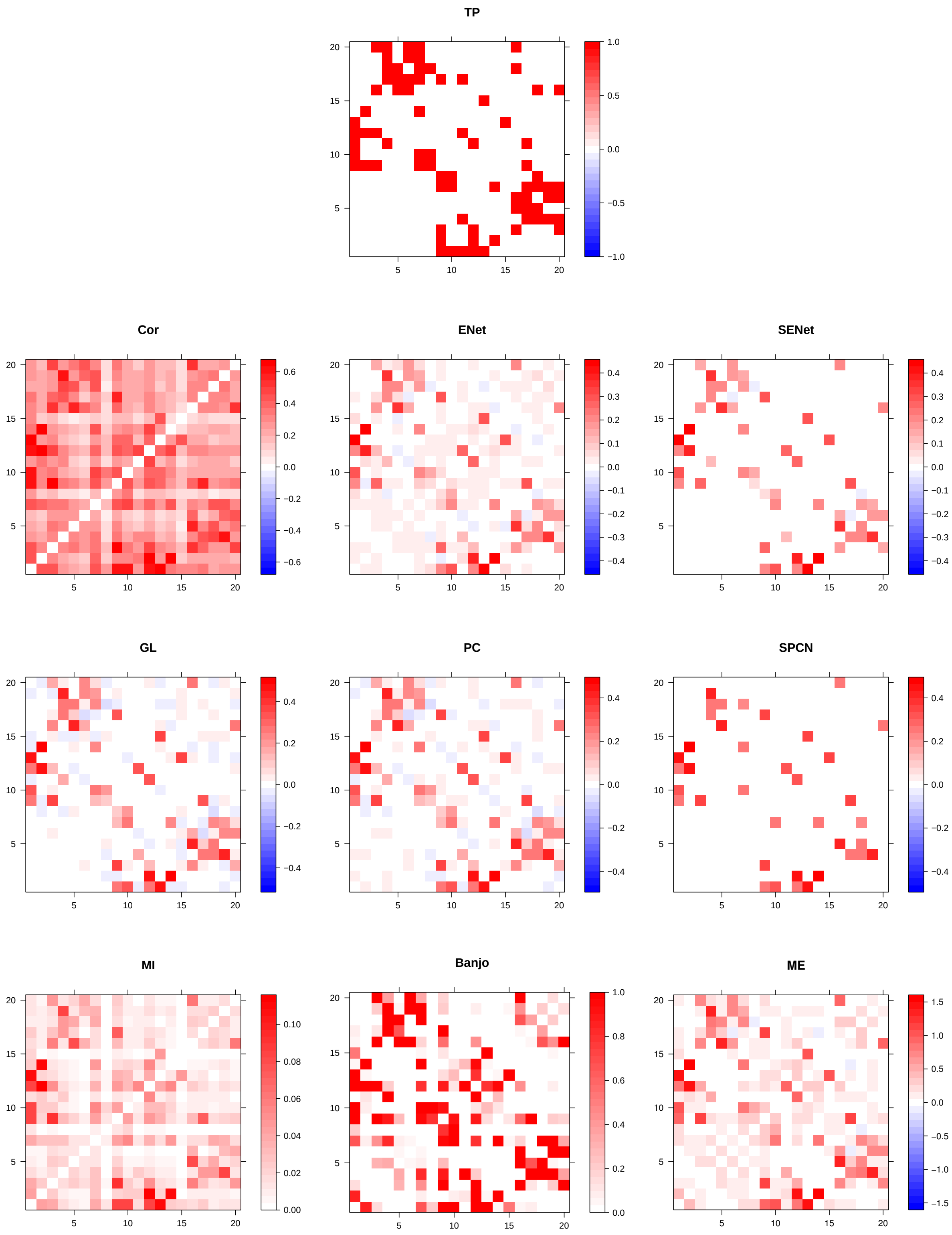

Figure 5.1.: *Heatmaps of the true simulated interaction network (TP) and the interaction scores for each possible edge obtained from the different network reconstruction methods on 5000 observations. The color indicates the magnitude of the interaction scores as indicated by the color key on the right of each heatmap.*

better performance of MI since the contrast between high and low scoring interactions is higher. A possible explanation might be the binarization of the data. The PC, GL, ME and ENet approaches show a high contrast between high and low scores. However, there are many scores close to zero. In the case of GL, ENet and ME, this might be improved by optimizing the penalization parameters. Comparing GL and PC, due to the Lasso regularization GL shrinks some of the low scores to zero but only in a few cases. The small scores might result from a large number of observations, giving support to the low scoring interactions. The results from Banjo also have a good contrast between the low- and high-scoring interactions. Banjo results in many high scoring interactions from which many are true interactions but some are also high-scoring false positives. This could potentially be improved by allowing more bootstrapping samples. When applying the cutoffs in SENet and SPCN as explained before, we recover much fewer interactions but only true ones. This reflects the objective of these two methods. Instead of delivering all possible interactions, they recover a few interactions that, however, have most support in data.

As an objective measure of the performance of the methods, we also compute the precision-recall (PR) and the receiver operating characteristics (ROC) curves. In the PR curves in Figure 5.2 the fraction of true positive predictions among all gold-standard positives (*Recall*) is compared to the fraction of true positive predictions among all positive predictions (*Precision*). The fractions are calculated along a sequence of cutoffs on the absolute interaction scores. An optimal method would quickly achieve a recall of one while keeping a high precision when lowering the cutoff. In the present example, this is achieved by GL and PC. The sparse methods SPCN and SENet detect around 50% and 75% of the true positives, respectively. The remaining interactions have a score of zero leading to a drop in the precision to the minimal precision as indicated by the straight lines. In contrast to mutual information and correlation these methods do not loose precision already at high cutoffs and thus increase the confidence in the detected interactions. ME, ENet, SGL and Banjo achieve a recall similar to SENet but then start loosing precision at low edge scores.

Figure 5.2 also shows the receiver operating characteristic curves and the corresponding area under the ROC curves, abbreviated by AUC. The ROC curves compare the recall, also called *sensitivity*, to the fraction of false positive predictions among all gold-standard negatives (*false positive rate* or 1-*specificity*). Hence, the higher the AUC the better the method performs. Again, SPCN has a low AUC that is explained by the same argument as described above: while lowering the cutoff many true positive predictions are added while not making any false positive predictions until none of the non-zero interaction scores is left. Note, that although MI and Cor achieve high AUCs (around 0.87), their performance in the precision-recall curves is low. Thus, while yielding relatively few false positive predictions out of all gold-standard negatives, MI and Cor have false positive predictions already at high cutoffs. The remaining methods have almost perfect AUC between $0.90 - 0.98$.

Since the analysis above might depend on the small exemplary network, we perform the analysis again on 25 networks with 50 nodes, 100 edges and 10.000 observations. The results averaged over the 25 networks are very similar to the results obtained in the small network (compare Figure 5.2 to Figure A.1 and A.2). Again, MI and Cor already loose precision at high cutoffs on the edge scores and but show AUCs similar to the other methods. SGL and SPCN have lower AUCs as compared to the other methods but all edges

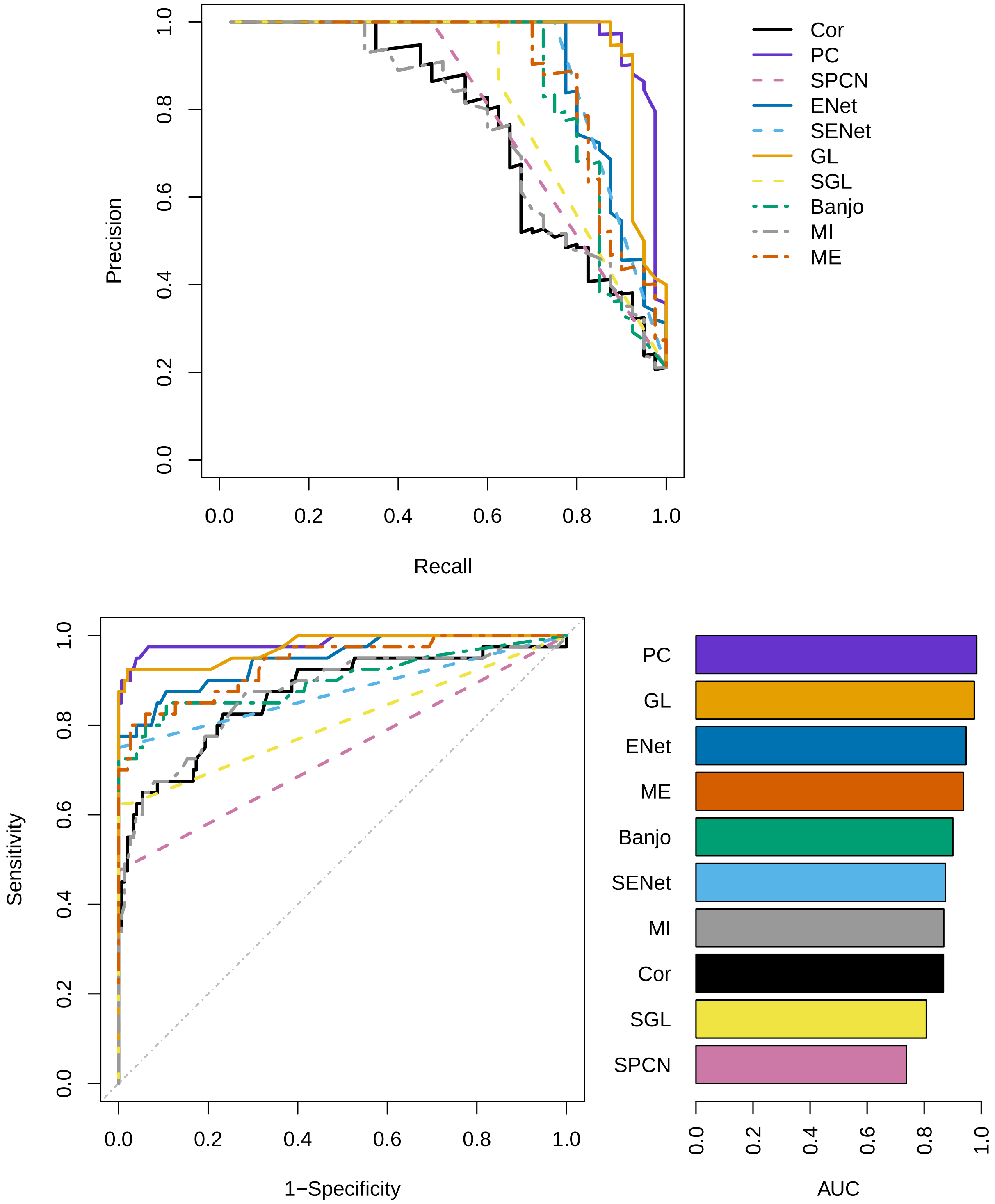

Figure 5.2.: *Precision-recall curves, receiver operating characteristics curves and the corresponding area under the curves (AUC) for a simulated network with 20 nodes and 40 interactions. The methods were run on 5000 observations. Solid and dashed lines indicate methods that are run on continuous data and their corresponding sparse version, respectively. Dotted-dashed lines indicate methods run on discrete data.*

having non-zero edge score are true positive predictions. There is almost no difference in the performance visible between the remaining methods.

5.1.4. Noisy data

There are different types of noise in the data possibly affecting the reconstruction of the networks, e.g. random noise affecting single observations in a sample or technical biases affecting certain observations over all samples. We check the robustness of the network reconstruction to these types of noise in the following section. we performed the analysis again on 25 networks with 50 nodes, 100 edges and 10.000 observations.

First, we add random Gaussian white noise, which is drawn independently for each observation in each sample, to a percentage of observations. This noise is simulated by a univariate Gaussian with a mean of 0 and a standard deviation of 1. Figure 5.3 shows the AUC over an increasing fraction of noise. All methods have relatively constant AUC over the different percentages of noise. SPCN and SGL perform worse than the other methods in the absence of noise. Upon increasing univariate noise these methods approach the AUC of the other methods. This is also visible in the PR curves (see Figure A.3). Since we observe this effect only for SGL and SPCN but not for PC or GL, it could result from the cross-validation scheme. The remaining methods, except MI and Cor, decrease slightly in their performance as indicated by the PR curves (see Figure A.3).

Next, we model biases affecting the whole observation vector. We use a multivariate Gaussian distribution with an arbitrarily chosen covariance matrix and a mean of 0 to model mutivariate Gaussian noise. The noise is added to a percentage of randomly chosen observations. This type of noise introduces a different covariance structure, which might affect the network reconstruction. The effect on the AUC of the multivariate noise is more pronounced compared to the univariate noise (Figure 5.3 and A.4). All methods except SGL loose AUC the more percent of noise is added to the data. Banjo and MI are most affected by increasing percentage of noise. This could be a result of the binarization, however ME is not affected.

5.1.5. Sample and network size

The network and the sample size can affect the robustness of the different network reconstruction methods. We thus compare the performance of the different methods with a variable number of nodes while keeping the sample size constant and *vice versa*. First, we generate 25 networks each consisting of 50 nodes and 100 edges. For each network we generate from 60 to 10.000 corresponding observations. Figure A.5 shows the average AUC for each number of simulated observations. As indicated by the average AUC, the performance improves with the number of observations available. The performance of the ME method with less than 1.000 observations might be improved by optimizing the regularization parameter for these settings. Figure 5.4 shows two exemplary plots of the PR curves at two different numbers of samples (100 and 10.000). Although, based on the AUC most methods perform better than random with only 60 observations, the PR curves indicate that the performance measured by the AUC is mostly boosted by the gold-standard negatives. In contrast, in the PR curves the improvement in the performance of the methods with more and more observations is much more pronounced (see examples in Figure 5.4 and Figure A.6).

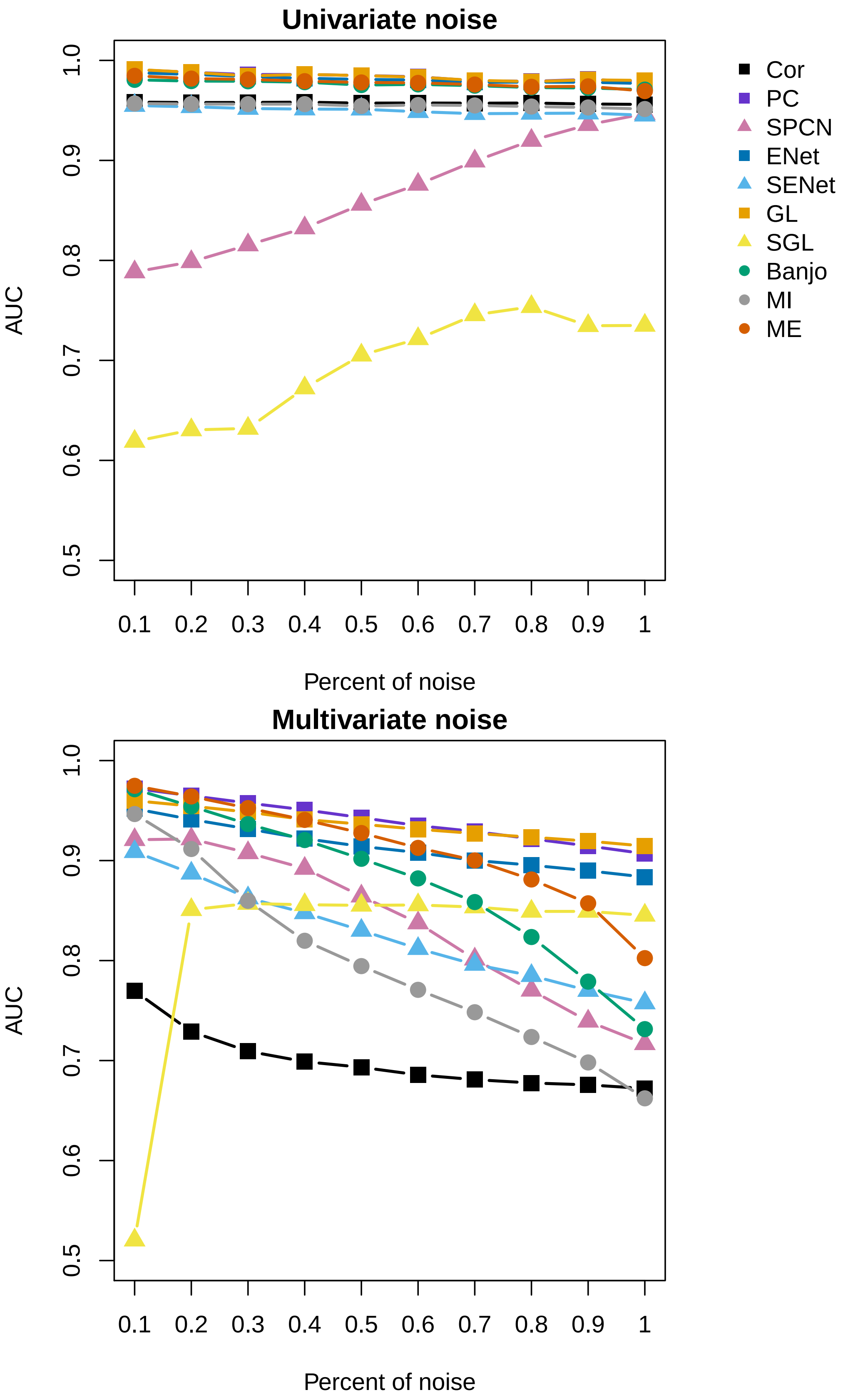

Figure 5.3.: *Average AUCs of the different methods over 25 simulated networks each having* 50 *nodes and* 100 *edges. Networks where inferred from* 10.000 *observations with an increasing percent of univariate and multivariate random noise.*

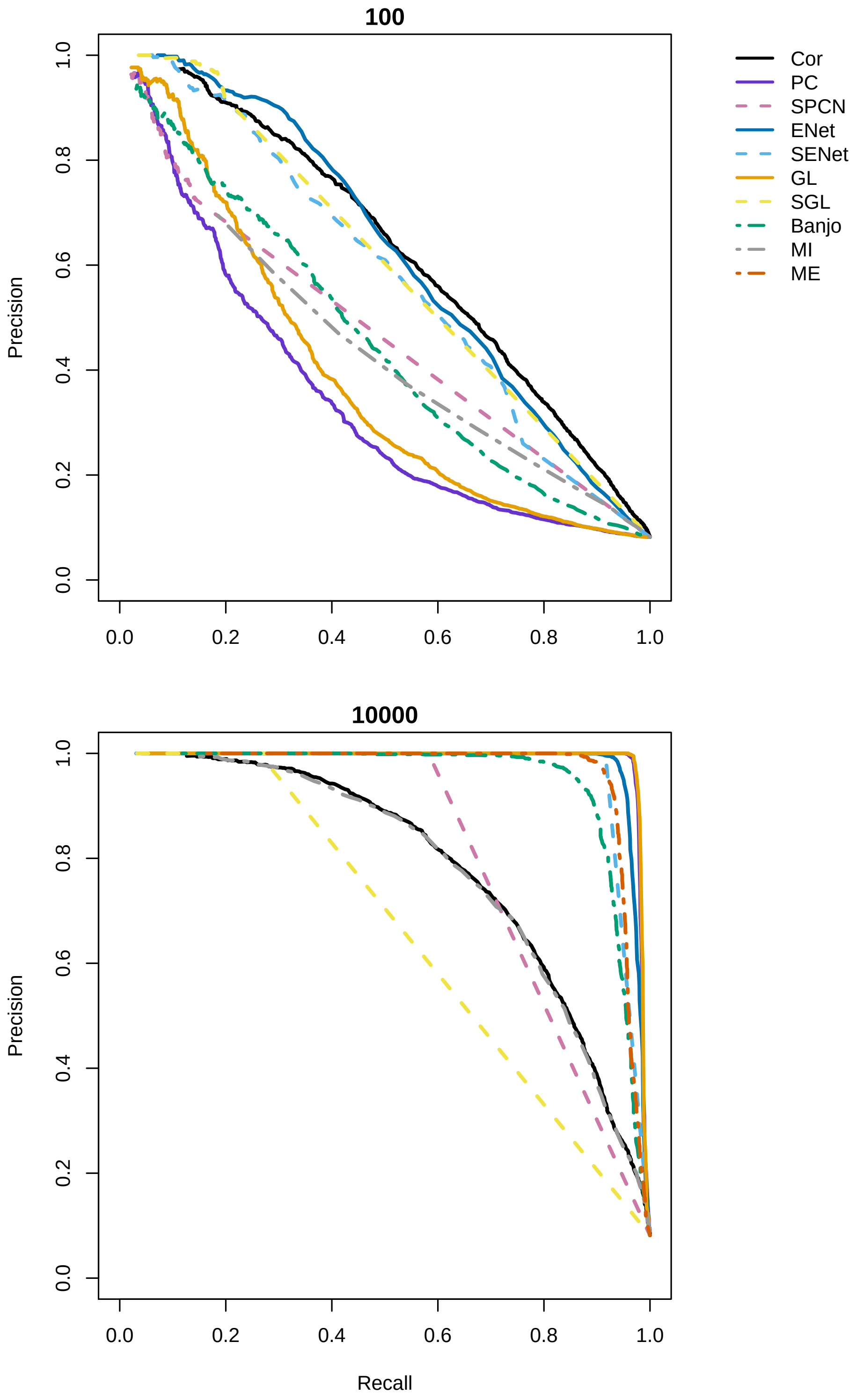

Figure 5.4.: *Two exemplary average precision recall plots for* 100 *and* 10.000 *observations, respectively. The performance of the different network reconstruction methods is averaged over* 25 *simulated networks each having* 50 *nodes and* 100 *edges. Solid and dashed lines indicate methods that are run on continuous data and their corresponding sparse version, respectively. Dotted-dashed lines indicate methods run on discrete data.*

Next, we investigate the behavior with an increasing number of variables while keeping the sample size constant at 10.000 observations. We increase the number of edges proportional to the number of variables, i.e. we keep the average node degree constant at 2. Some methods did not terminate within the time frame given when being tested with more than 500 variables, which is why we restricted the comparison to between 10 and 100 variables. As shown in Figure A.7, the PR curves improve for most methods (except SGL) with an increasing number of variables when the number of edges is kept constant. The average AUC stays more or less constant for all methods over for a increasing number of variables (see Figure A.8).

5.1.6. Graph density

The network reconstruction methods enforce sparseness on the edges. In very dense graphs this might lead to high false negative rates. However, in sparse networks this should help recover the true interactions. Here, we investigate the performance of the different methods with respect to the graph density. We use a simulated graph with 50 variables and vary the number of edges in the true network between 100 and 900. The number of observations is kept constant at 10.000.

Most methods perform best for 100 edges based on the average AUC over 25 networks (see Figure A.9). The AUCs for PC, GL and ENet stay more or less constant at a high level over the various numbers of edges. Also the AUCs for SPCN and the SGL do not change over the different number of edges. However, their performance is around 0.7 and 0.5, respectively. The remaining methods including ME, SENet, Banjo, MI and Cor drop slightly in their performance the more edges are present. This can also be observed in the two exemplary plots of the PR curves (Figure 5.5, see also Figure A.10) for 100 and 900 edges, respectively. While for the low number of edges Banjo performs similarly to the other methods, its performance drops to similar levels as the Cor and MI approach. The same but less pronounced can be observed for ME, whose performance is similar to SPCN in the case of very dense graphs.

5.1.7. Conclusion

In summary, based on our analysis there are only slight differences between the performance of the different methods under the scenarios tested on simulated data. The biggest difference in performance is observed by comparing Cor and MI to the remaining methods. As expected, Cor and MI detect many false positive edges and result in high edge scores for these edges. In contrast, the remaining methods result in good performances when tested on simulated data with varying number of variables and number of observations. Also with small percentages of noise, as simulated by random univariate Gaussian noise, all methods perform equally well.

Most pronounced are the conceptual differences between the sparse methods and the remaining methods. Under all conditions tested the sparse methods, especially SPCN, are very robust. Only, when increasing the graph density the performance of SENet drops. However, by adjusting the cutoff this issue might be resolved. On the other hand, in beneficial settings, e.g. when a large number of observations is available, the sparse methods might be more conservative than the other methods.

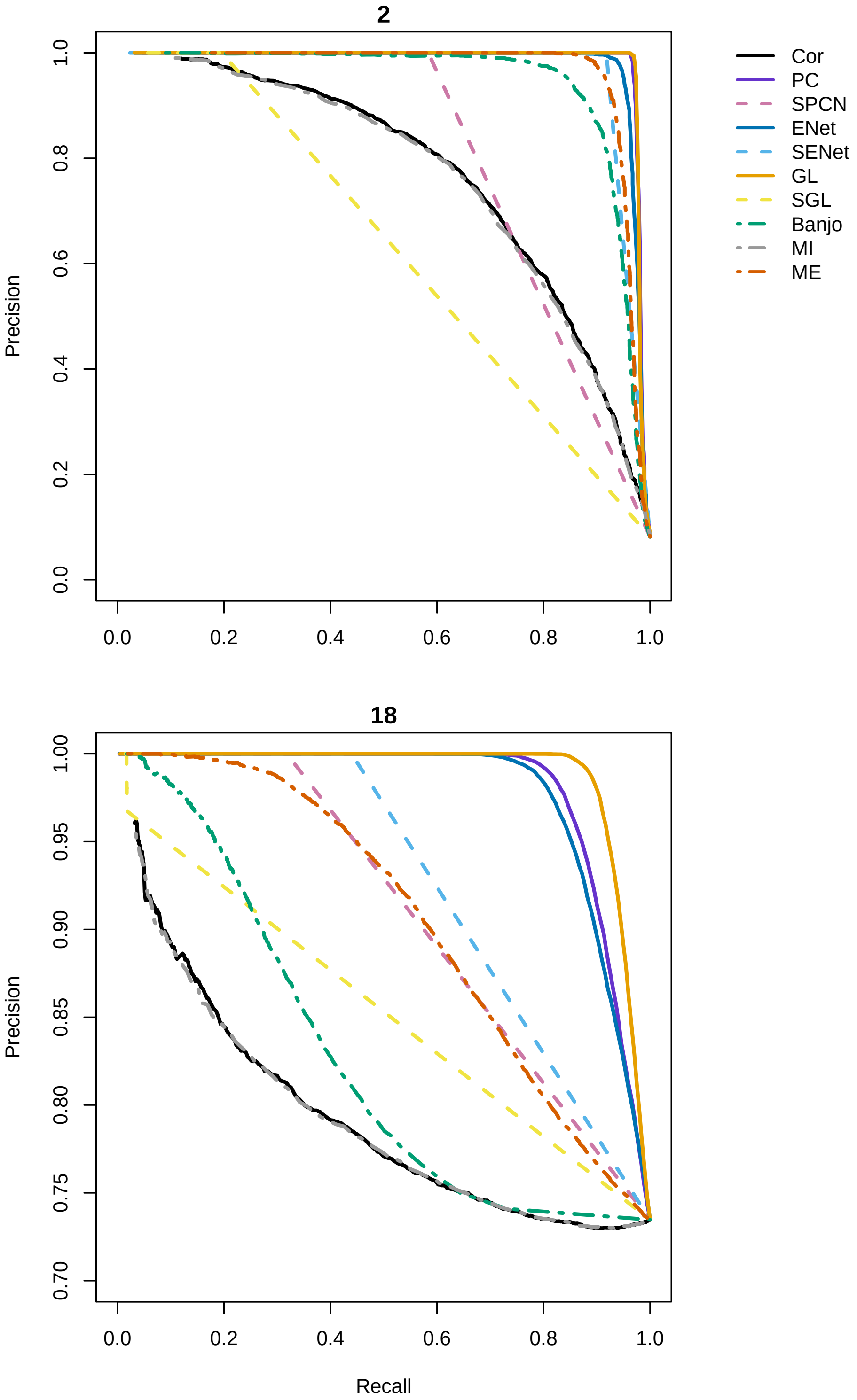

Figure 5.5.: *Two average precision recall plots for* 100 *(node degree* 2*) and* 900 *(node degree* 18*) edges, respectively. The performance of the different network reconstruction methods is averaged over* 25 *simulated networks each having* 50 *nodes and based on* 10.000 *observations. Solid and dashed lines indicate methods that are run on continuous data and their corresponding sparse version, respectively. Dotted-dashed lines indicate methods run on discrete data.*

5.2. A method comparison on experimental data

In a recent study [53] the authors used the DamID technique to produce a comprehensive data set of known and unknown chromatin-binding proteins. This data set comprises 179.548 genomic locations and 112 samples. The set of samples contains 37 proteins that were previously unknown for binding chromatin, 70 known chromatin-binding proteins and ChIP-Chip data for 5 histone marks. To the best of our knowledge this is the largest uniformly processed data set publicly available. In the following section, we use this data to compare the results of the different network reconstruction approaches on real biological data.

5.2.1. DamID data for Drosophila

In [53] the authors performed DamID experiments (see Section 2.3.3) followed by a genomic tiling array analysis to identify novel chromatin-binding proteins in Drosophila. We downloaded [1] the fully pre-processed, loess-normalized and median-centered data set and the discretized data set for the *Kc167* cell line. As explained in [53], the authors discretized each sample into enriched, intermediate and depleted domains using a 3-state Hidden Markov Model (HMM). Only the enriched state was called as target while the rest was defined as non-targets. In general, most of the samples show little or no enrichment in most of the regions and for the discrete data on average 10% of the regions are called as target sites (see Figure A.11).

5.2.2. Methods

For the method comparison on the Drosophila data set we run the following methods. We use the correlation-based, the (Sparse) Elastic Nets and the (Sparse) Graphical lasso approach on the continuous data. The partial correlation and SPCN are computed on ranked data. We use the discretized data to run the mutual information and the maximum-entropy approach. The results for Banjo are obtained from [53].

5.2.3. Edge Rankings

Most of the methods, except SPCN, SENet and Banjo, return a low percentage of edges with scores equal to zero (see Table 5.2 column "%E(w=0)"). Usually a cutoff is chosen to retain the most important interactions for visualization. However, as the choice of the cutoff is more or less arbitrary, we first compare the correlation computed between the ranked absolute edge scores per method.

The rank correlations between the edge scores of the different methods range from around 0.08 to 0.97 (see Figure 5.6). The highest correlation is obtained between the ENet and GL, which is expected since the underlying mathematical principles are very similar (see Section 3). Further, also the other covariance selection methods, except SGL, yield relatively similar results. SGL gives very few non-zero edges and thus is not correlated with any of the other methods. This is most probably due to the masking applied to unstable edges (see Section 5.1.2) suggesting that the results obtained by GL are not

[1] GEO database accession number *GSE36175*; June 25th, 2013

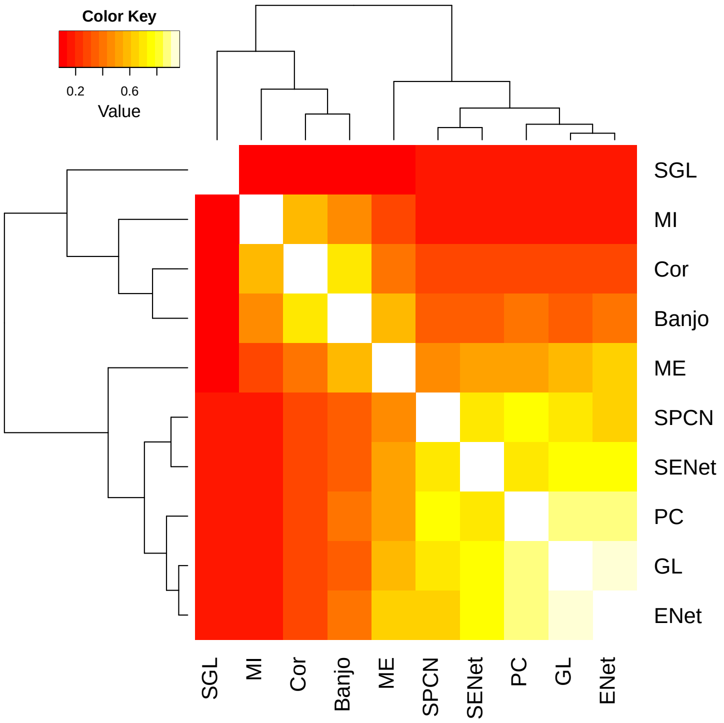

Figure 5.6.: *The heatmap shows the correlation coefficients computed on the ranks of the absolute edge scores resulting from each network reconstruction method on the Drosophila data set. Red cells correspond to low correlation while yellow indicates high correlation as indicated in the color key.*

robust. Interestingly, Banjo is more correlated with MI and Cor, while ME is correlated with Banjo and additionally with the group of covariance selection methods.

5.2.4. Network features

Since many methods have a high percentage of non-zero scores (see Table 5.2), we need to apply a cutoff on the edge scores to compare different features of the resulting networks. For the remainder of this section, we call an interaction between two nodes if the scores returned by each method exceeds a certain threshold (see column “Cutoff” in Table 5.2). We choose the cutoffs arbitrarily by looking at the distribution of edge scores (see Figure A.12). For the SPCN and the sparse Elastic Nets approach we require a non-zero score allowing about 25% of the possible edges. For the Banjo approach we use the cutoff given in the original publication [53] resulting in a very low number of interactions. As seen in Section 5.1.3, PC, ENet and GL have many edge scores that only slightly deviate from zero. The same is true for the edge scores on the *Drosophila* data set, which is why we choose a cutoff on the absolute edge scores slightly above zero.
Applying the cutoffs to the networks yields very different numbers of edges and nodes. which take part in at least one edge, for the different network reconstruction methods. The highest numbers of interactions are achieved by Cor, SPCN and SENet. This includes two sparse networks for which we only required an edges score different from zero. Banjo has the lowest number of edges due to the stringent choice of the cutoff. PC, GL and ME have similar numbers of edges around 800. The resulting networks connect (almost)

all nodes for the following methods: PC, SPCN, SENet and ME. SGL connects less than half of the nodes. The global cutoff choice leads to several unconnected nodes for Cor, ENet, SGL, Banjo and MI. Further, for most methods, except SGL, Banjo and ENet, the average node degree lies between 14 and 31.

Table 5.2.: *The table compares different network features of the reconstructed Drosophila network for each of the network reconstruction method. The first column indicates the percent of edges with a edge scores of zero. The second column gives the arbitrarily chosen cutoff on the edge scores to obtain the top-scoring interactions. The following two columns give the resulting numbers of edges and of connected nodes, respectively. Lastly, the remaining columns show the average, maximum and minimum node degree, respectively.*

	% E(w=0)	Cutoff	#Edges	#Nodes	Avg. ND	Max. ND	Min. ND
Cor	0.00	0.50	1529	102	27.30	64	0
PC	0.13	0.05	709	112	12.66	25	1
SPCN	75.71	0.00	1510	112	26.96	49	12
ENet	0.85	0.10	473	110	8.45	40	0
SENet	72.18	0.00	1729	112	30.88	61	11
GL	0.16	0.05	979	111	17.48	37	0
SGL	98.50	0.00	93	45	1.66	23	0
Banjo	44.48	0.70	141	98	2.52	8	0
MI	4.50	0.05	1095	89	19.55	57	0
ME	0.58	0.50	829	112	14.80	57	1

5.2.5. Known protein-protein interactions

Evaluating the performance on this biological data set is difficult because there is no gold-standard set of interactions available. However, as the goal is to learn the biological function of the proteins from co-localization, we hope to detect many known protein-protein interactions (PPI). There are many aspects that have to be taken into account when interpreting the comparison to PPIs. First, PPIs are conceptually different from co-localization and hence we expect to detect many interactions that are not present in the PPI database. Second, PPI databases are far from complete, e.g. most databases do not contain information on interactions involving specific post-translational modifications of the proteins. Third, the PPIs might contain false positives or indirect interactions, rendering the interpretation of the performance measures difficult. Lastly, these databases rarely contain information about true non-existing interactions, i.e. most interactions are just not studied or it is hard to conclusively rule out that an interaction exists. Despite these drawbacks, PPI databases are probably the most useful information resource that can be obtained for evaluating the network reconstruction methods.
Here, we compare the resulting networks to known PPIs from the DroID-database [66] and manually selected interactions from [55] that overlap with the data set. From the DroID-database [2] we include the interactions from different high-throughput yeast-to-hybrid and co-affinity-purification screens. Further, we include the physical interactions retrieved from other databases such as BioGrid and MINT. Overall, we retrieved 74 known interactions between the proteins in our network out of the 6.216 interactions from the database.

[2] Downloaded August 11th, 2014

Since there are only 74 known interactions the interpretation of the PR curves and the ROC curves using these PPI-derived gold-standard set (Figure 5.7) is difficult. The best performing methods based on the AUC are Banjo and ENet. However, the difference between the various methods is not very pronounced with AUCs ranging from 0.53 to 0.67. Judging by the PR curves, the best performing methods include GL, SGL, SENet and ENet, as well as SPCN and Banjo. However, overall the performance of the methods is rather similar.

We also investigate the overlap of the top-scoring edges, selected as in Section 5.2.4, to the gold-standard positive interactions. The biggest overlap is achieved by SENet, SPCN and Cor. The recall of the different methods are at comparable levels. However, based on the precision, SENet, SPCN and Cor perform best.

Table 5.3.: *Number of overlapping edges, Precision and Recall calculated by comparing the gold-standard interactions with the top-scoring interactions, as selected in Section 5.2.4, for each method.*

	Cor	**PC**	**SPCN**	**ENet**	**SENet**	**GL**	**SGL**	**Banjo**	**MI**	**ME**
# Overlap	30.00	21.00	31.00	16.00	35.00	28.00	5.00	11.00	21.00	21.00
Precision	0.41	0.28	0.42	0.22	0.47	0.38	0.07	0.15	0.28	0.28
Recall	0.02	0.03	0.02	0.03	0.02	0.03	0.05	0.08	0.02	0.03

5.3. Summary and discussion

In summary, we have compared the different network reconstruction methods described in Chapter 3 on simulated and biological data. With the simulated data, we could compare the methods under different settings and calculate the performance based on the true underlying network. The different settings studied here include increasing numbers of variables, edges and observations, as well as different percentage of univariate and multivariate noise that are intended to simulate different sources of technical noise. The resulting networks from biological data were compared to each other and to a PPI database. From our analysis, we conclude that the different methods perform equally good on the simulated and the biological data. Our analysis highlights the conceptual differences between the sparse methods (SPCN, SENet and SGL), the pairwise association methods (MI and Cor) and the remaining methods. With the edge scores of the pairwise association methods it is very difficult to distinguish between true interactions and false positives. The sparse methods tend to make very few false positive prediction but at the same time sacrifice some true positive predictions.

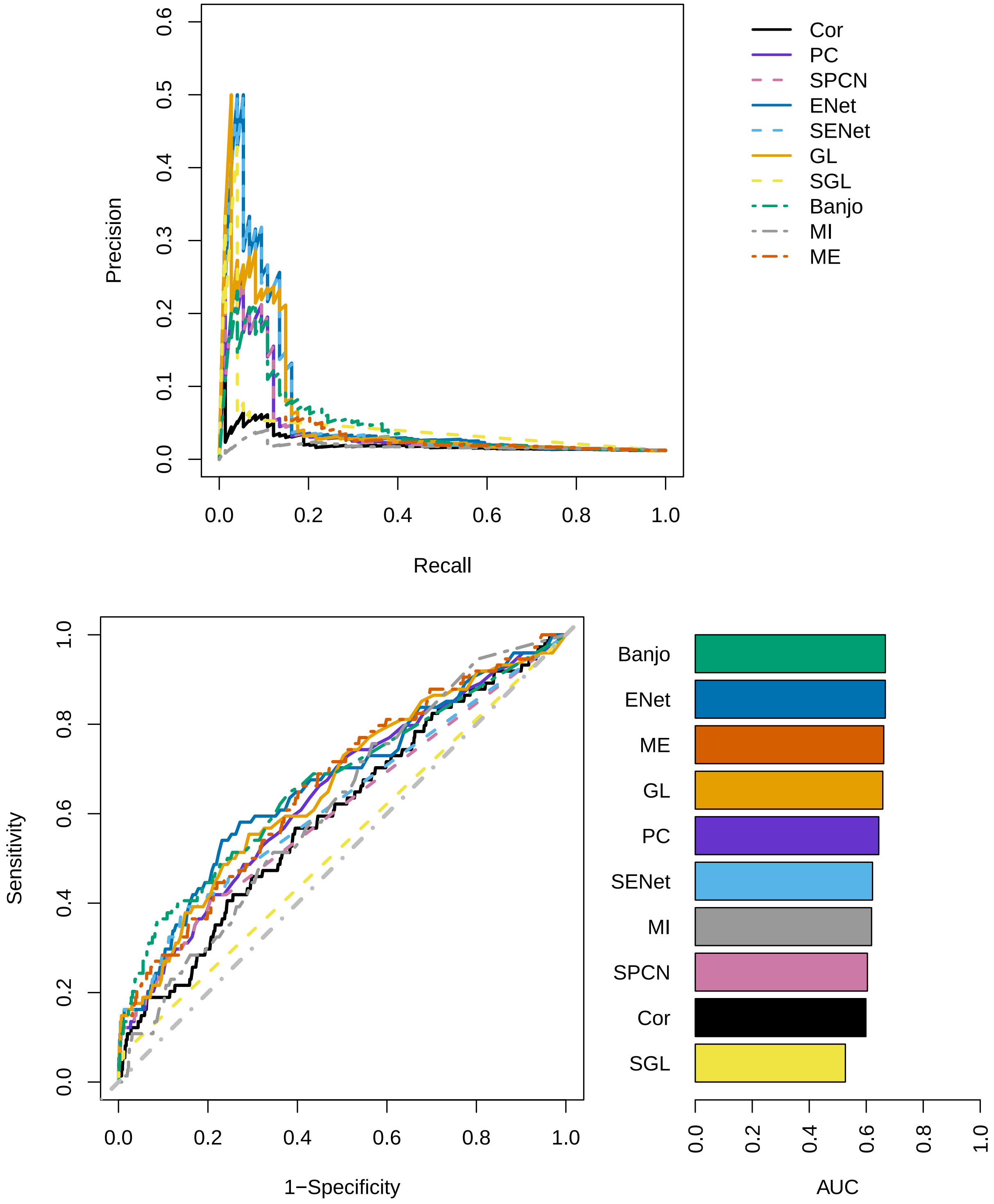

Figure 5.7.: *Precision-recall, ROC curves and the corresponding AUCs of the different methods on the Drosophila data set. Only 74 gold-standard positive interactions connecting nodes in the Drosophila data set were available in the DroID-database. Solid and dashed lines indicate methods that are run on continuous data and their corresponding sparse version, respectively. Dotted-dashed lines indicate methods run on discrete data.*

CHAPTER 6

The human chromatin signaling network at promoters

Chromatin is regulated by a complex signaling network consisting of interactions between histone modifications and chromatin modifiers. These interactions are often unknown, based on the analysis of a single or a few genes, or studied *in vitro*. Several recent studies [53, 55] have used genome-wide DNA-protein binding data to reconstruct chromatin signaling in Drosophila. We have compared the different approaches on Drosophila data in the previous chapter. In the following chapter, we recover interactions between chromatin modifiers and histone modifications from genome-wide ChIP-Seq data at promoters in human cells. To the best of our knowledge this is the first chromatin signaling network in humans. Many of the interactions discovered from co-localization have biochemical support in the literature but also unknown interactions are revealed. These novel interactions can provide the basis for mechanistic hypotheses on chromatin signaling, of which we discuss several examples in the light of existing literature. Parts of the following chapter are published in [1].

6.1. Data preparation

Over the past years many large-scale projects performed experiments or computational analyses to reveal the regulatory state of human cells. The ENCODE [26] and similar consortia produced a large amount of ChIP-Seq data sets including data on HMs, CMs and TFs. The coding regions of the genome are annotated and collected in, for example, the RefSeq database [67]. We use these large collections of data and standardized annotations to retrieve normalized ChIP-Seq levels, which are then used to explore the chromatin signaling at human promoter regions. The following section describes the data retrieval and normalization.

6.1.1. Promoter annotation

A common practice in retrieving promoter annotations is to define a region around the 5'end of a known transcript as promoter. Here, we define the broad promoter region as a 4.000*bp*-window centered at the TSS of each annotated RefSeq gene. We remove redundant transcript annotations and obtain 26.893 promoter regions. Note, that these broad promoter regions might also capture signals from the (transcribed) 5'-part of the genes and enhancers located in proximity of the genes. However, this large region also allows to detect relationships that appear between the core promoter and the promoter-proximal region.

6.1.2. ChIP-Seq read counts in human cells

For the following analysis we use mostly data from the K562 cell line but also from the H1 human embryonic stem cell (hESC) that is publicly available (summarized in Table B.1). A convenient feature of this data set is that most of the experiments were performed in the same laboratory and thus we are likely to circumvent technical laboratory-dependent artifacts. Using Bowtie [68], we re-mapped the uniquely mapping reads to the *hg19* genome in a uniform way, to ensure that there is no bias resulting from different mapping techniques. We counted for each sample the number of ChIP-Seq reads that fall into the promoter regions using Rsamtools, thereby summing up replicates if available.

Read count normalization

ChIP-Seq experiments show technical artifacts resulting from the enrichment of DNA regions that are easy to pull down in the experiment but not necessarily bound by the protein of interest. Figure 6.1 shows the read counts of a sample plotted against the input DNA, i.e. the DNA that can be pulled down without the use of a specific antibody. A clear dependency of the sample reads on the input becomes apparent. The promoters bound specifically by the protein of interest, are visible as a diffuse cloud in the upper part of the plot. While the remaining promoters with no protein bound resolve as a compact cloud of points showing strong correlation with the input signal.
A common practice to account for such technical artifacts is to normalize the data against the input DNA. We apply a normalization technique that estimates the median increase in the signal against the input by equation 6.1. This formula estimates the slope of the correlation between the read counts of the sample S against the read counts of the input C. For both S and C we add a pseudo-count of 1 to avoid zero counts.

$$m = median(\frac{S+1}{C+1}) \tag{6.1}$$

We then define a normalized score S_{norm} by replacing S with the enrichment of the sample over the median-amplified input:

$$S_{norm} = \frac{S+1}{(C+1)*m} \tag{6.2}$$

A convenient feature of this normalization method is that it shrinks the read counts that can be explained by the input towards zero but, in contrast to a linear regression of the sample against the input, avoids non-negative values (see Figure 6.1). The latter is

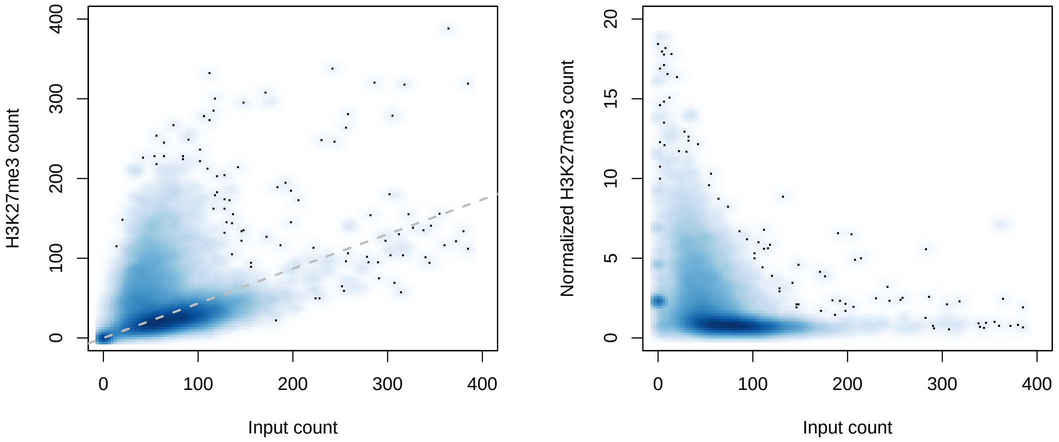

Figure 6.1.: *Scatterplot of the ChIP-seq signal before and after normalization, respectively, against the input signal. The dashed line indicates the median fold-change as function of the input signal.*

important since negative read counts are counter-intuitive and hard to interpret in the context of ChIP-Seq experiments. However, it should be noted that this procedure relies on the assumption that most of the regions are not bound by a protein. The normalization depends on the clear enrichment of the sample over the input, i.e. the slope is hard to estimate if there is no clear signal in the sample or if all regions show enrichment. Finally, for technical reasons the normalized read counts were log-transformed and scaled to have mean zero and standard deviation one. The final distributions of the read counts by samples in K562 are shown in Figure A.13.

Gene expression data

To measure transcriptional initiation and gene expression we used CAGE data. This data was downloaded fully preprocessed from the UCSC genome browser[1]. We counted the reads in each promoter region and averaged replicates. The read counts were log-transformed and scaled to have mean zero and standard deviation one.

6.2. Chromatin signaling and gene expression

The presence of certain HMs and CMs at the promoter or the gene body is linked to the transcriptional state of the corresponding gene [36, 37]. Moreover, the HM levels at promoters can even quantitatively predict the steady state levels of mRNAs [69–71]. In the following section, we check how much information about gene expression is contained in the available set of HMs and CMs. Furthermore, we investigate whether these HMs and CMs are redundant predictors of gene expression and finally, whether this trend is preserved across cell types.

[1] Sample names: K562CellPapAlnRep1/2.bam and H1hesc CellPapAlnRep1/2.bam (Nov. 2012)

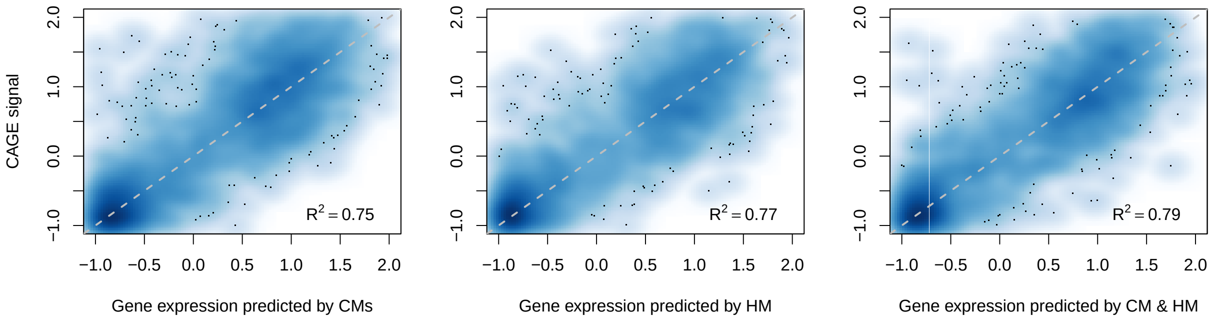

Figure 6.2.: *Predicted expression levels plotted against the measured CAGE tag counts in the K562 cell line. Expression was predicted using linear regression in a* 10*-fold CV from three different sets of variables: only histone modifications, only chromatin modifiers or both.*

6.2.1. Predicting gene expression

HMs and CMs are part of a chromatin-signaling network connected to gene expression. As such both, HMs and CMs, should contain information about and be predictive of gene expression. We measure how informative the available HMs and CMs are by using linear regression and predicting the gene expression values by the corresponding HM or CM levels in the promoter of the genes. As shown in Figure 6.2, the HM levels explain about 77% and the CM levels explain 75% of the variance in gene expression. The results for HMs are similar to previous work [69–71]. The good predictive performance confirms that both HMs and CMs contain information about gene expression.

6.2.2. Predicting histone modifications and chromatin modifiers

Next, we test whether the HMs and the CMs are coupled in the same chromatin-signaling network. If so, using both, HMs and CMs, as predictors in the linear regression should not add more information on the expression values. By combining HMs and CMs we improve the predictions only by 3% or 4% compared to using only HMs or CMs, respectively. This rather small increase in the predictive power suggests that the available HMs and CMs hold redundant information about gene expression and thus might be together involved in the same chromatin-signaling network that regulates gene expression.

Further, we checked if the HMs and the CMs at hand interact in the same chromatin-signaling pathway. If so, they should be predictive for each other. This hypothesis can be tested by predicting HMs from CMs and *vice versa*. The resulting R^2-values are shown in Figure 6.3. For each HM, the CMs account for at least 50% of the variance. In predicting the CM levels from HM levels, the models account for over 50% of the variance (see Figure 6.3) for most of the CMs. Thus, for those well-predicted CMs the HMs in the data set might cover most of the recruitment mechanisms or enzymatic targets. From the high predictive power of the models we conclude that the set of HMs and CMs are both involved in the same chromatin-signaling pathway affecting gene expression.

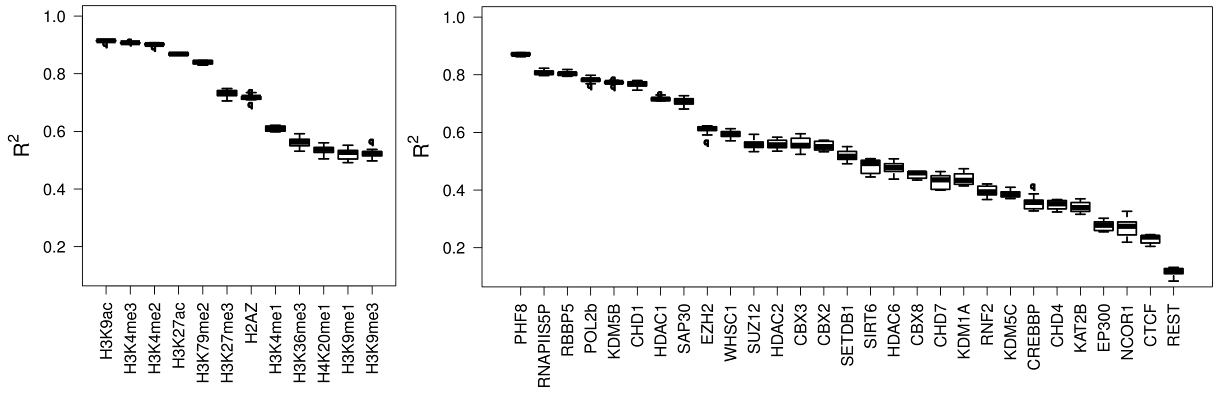

Figure 6.3.: *Predicting histone modifications from chromatin modifiers and vice versa in the K562 cell line using linear regression in a* 10*-fold CV. Each box summarizes the test-R^2 distribution over the* 10 *CV-folds.*

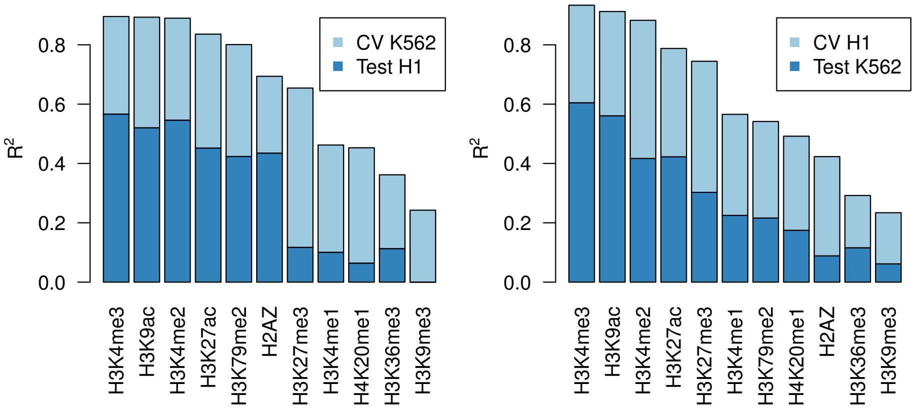

Figure 6.4.: *Predicting histone modifications from chromatin modifiers across cell types. First, the models are trained in the human K562 cell line and then tested in the H1 ESCs. Second, the models are trained in the H1 ESCs and tested in the K562 cell line. Light color indicates the R^2-values obtained on the CV-test set and dark color indicates the R^2-values in the other cell type.*

6.2.3. Conservation across cell types

Under the assumption that the HM-CM-interactions reflect common functions of the proteins in transcription and its regulation, we expect that the contribution of a CM to the prediction of an HM in one cell type is similar in another cell type. Thus, given the regression model trained on the data from the K562 cells we should be able to predict the HM levels in another cell type. We tested this using ChIP-Seq data for the 14 CMs and the 11 HMs in hESCs that were also measured in the K562 cells. Indeed, the regression models learned from the data available for both cell types show good agreement (Figure 6.4). The lower performance of the models when tested on the data from a different cell type is expected due to biological variation, e.g. different expression levels of the CMs. Overall, the quantitative effects of the interactions within the chromatin-signaling network are preserved suggesting cell-type independent interactions involved in transcription and its regulation.

6.3. Human chromatin-signaling networks

In the previous section we have shown that the HM and CM levels at the promoter regions are predictive for gene expression of the associated gene. Further, we argued that HMs and CMs are jointly involved in signaling-pathways affecting gene expression. In the following section, we investigate the interactions between these CMs and HMs at different levels of details: from correlated binding profiles to high-confidence pair-wise interactions. We will focus on interactions between HMs and CMs, assuming that HMs and CMs act on different layers in the signaling pathway. These interactions allow us to connect the very localized, 'passive' signal of the HMs with the CMs, that in turn are able to actively remove, interpret and maintain the HMs. The analysis will shed light on the function of individual HMs in the chromatin-signaling.

6.3.1. Co-localization patterns

We first investigate the global co-localization of the HMs and CMs at promoters. The complete correlation matrix is shown in Figure A.14. However, comparing the correlations between CMs to the original publication [4], we have similar results and thus do not discuss these correlations in more detail. Figure 6.5 shows the pairwise Pearson correlation coefficients between each pair of HMs and CMs only. A clear separation between three groups of CMs correlated with different sets of HMs become apparent suggesting differential function of these groups of CMs. The first group of CMs, including CBX8, CBX2, SUZ12, EZH2, CREBBP and CBX3, is related to HMs that are known to be involved in gene repression (i.e. H3K27me3, H3K9me3 and H3K9me1, as well as H4K20me1). The second group has no apparent specificity for any HMs. Finally, the last group of CMs (i.e. HDAC1, PHF8, RNAPIIS5P, PolII, CHD1, SAP30, KDM5B and RBBP5) is highly correlated with HMs that are known to be present at active promoters (H3K4me3, H3K4me2, H3K79me2, H3K9ac and H3K27ac). Indeed, the last group of CMs and HMs achieves the highest correlation coefficients, however no individual interactions become apparent. Overall, both correlation matrices (Figure 6.5 and A.14) show groups of highly correlated HMs and CMs. The HMs and CMs within such a group are mostly strongly anti-correlated with the HMs and CMs of another group. However, within the groups of HMs and CMs the correlation coefficients are very high rendering the understanding of the individual interactions difficult.

6.3.2. A coarse grained interaction network

After investigating the raw correlation signals between HMs and CMs, we next investigate which interactions are most specific for each HM. The goal of this analysis is to get a global overview of the interaction network that might already reveal pathways connecting different HMs and chromatin states. For this we used the Sparse Elastic Net approach described in section 5.1.2 to predict each HM by the CMs. The Elastic Net assigns high coefficients to only those CMs that are most informative for the prediction of an HM. Thereby similarly good predictors, which would allow for the detection of, for example, protein complexes that interact with an HM, obtain similar weights. Note that due to the separation of HMs and CMs this approach does only account for correlations within the CMs but does not take into account correlations among the HMs. As a consequence similar predictors might not only be protein complex members but also CMs connected

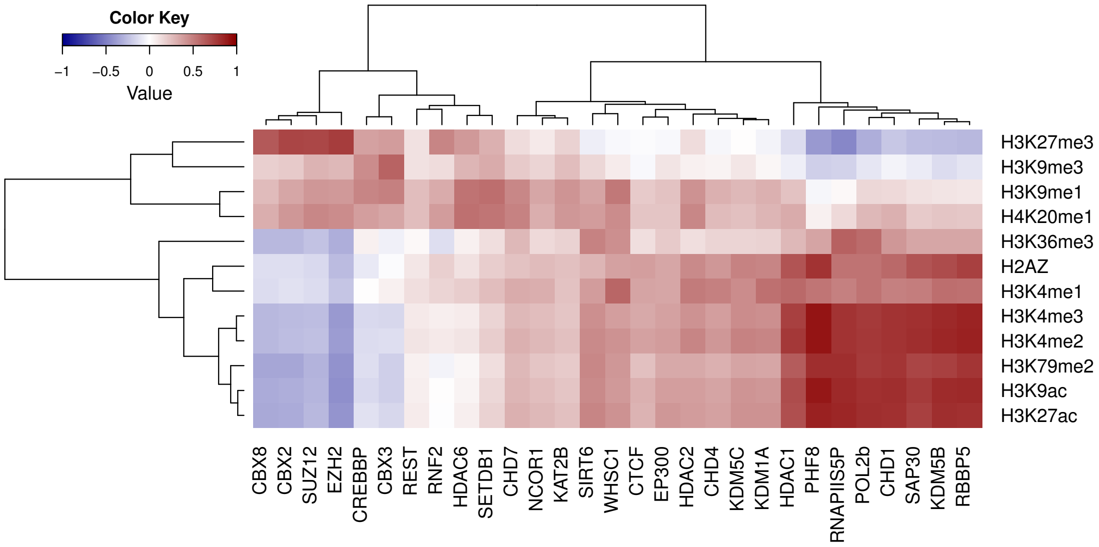

Figure 6.5.: *A heatmap of the Pearson correlation coefficient as indicated by the color key between histone modifications and chromatin modifiers.*

to a highly correlated HMs, leading to similar sets of CMs as predictors of these HMs. However, as this approach is less stringent than the other sparse methods it might reveal global patterns that help in understanding the complex behavior of the systems.

Figure 6.6 shows the resulting Sparse Elastic Net network. Each HM is linked to a different set of CMs indicating the different specificities of the CMs towards the individual HMs. As expected, we see a group of highly connected CMs and HMs involving the HMs H3K79me2, H3K4me3, H3K9ac and H3K27ac, which are highly correlated (see Figure A.14). However, we see also many interactions connecting the CMs to other HMs that are not directly related with active transcription. This suggest that these CMs might also be involved in pathways that communicate between different regulatory states of promoters. For example, the mutually exclusivity of PHF8 and H4K20me1 or H3K9me1 reflected as a negative interaction in the Figure 6.6 fits to the known function of PHF8 as a demethylase of H4K20me1 and H3K9me1 and might be necessary to block a repressive chromatin state at active promoters.

6.3.3. Deriving a high-confidence interaction network

We next turn from detecting pathways and groups to resolving the network structure towards the most specific interactions of each HM. We apply SPCN to all available HMs and CMs. As explained in section 5, SPCN is very conservative in detecting interactions. However, the failure to recover an interaction should not be interpreted as the absence of a biologically meaningful interaction. To increase our confidence in a recovered edge, we intersect the edges from the Sparse Elastic Net approach as described in section 6.3.2 with the SPCN. Hence, we retrieve a network that consists only of interactions recovered by both methods. With this methodology we first choose those CMs that are consistently highly predictive for an HM and then disregard those interactions that may have been induced by high correlations in the HMs by the SPCN.

Figure 6.7 shows the resulting interaction network between the HMs and the CMs. The strict selection of the SPCN method becomes most apparent on the cluster of HMs de-

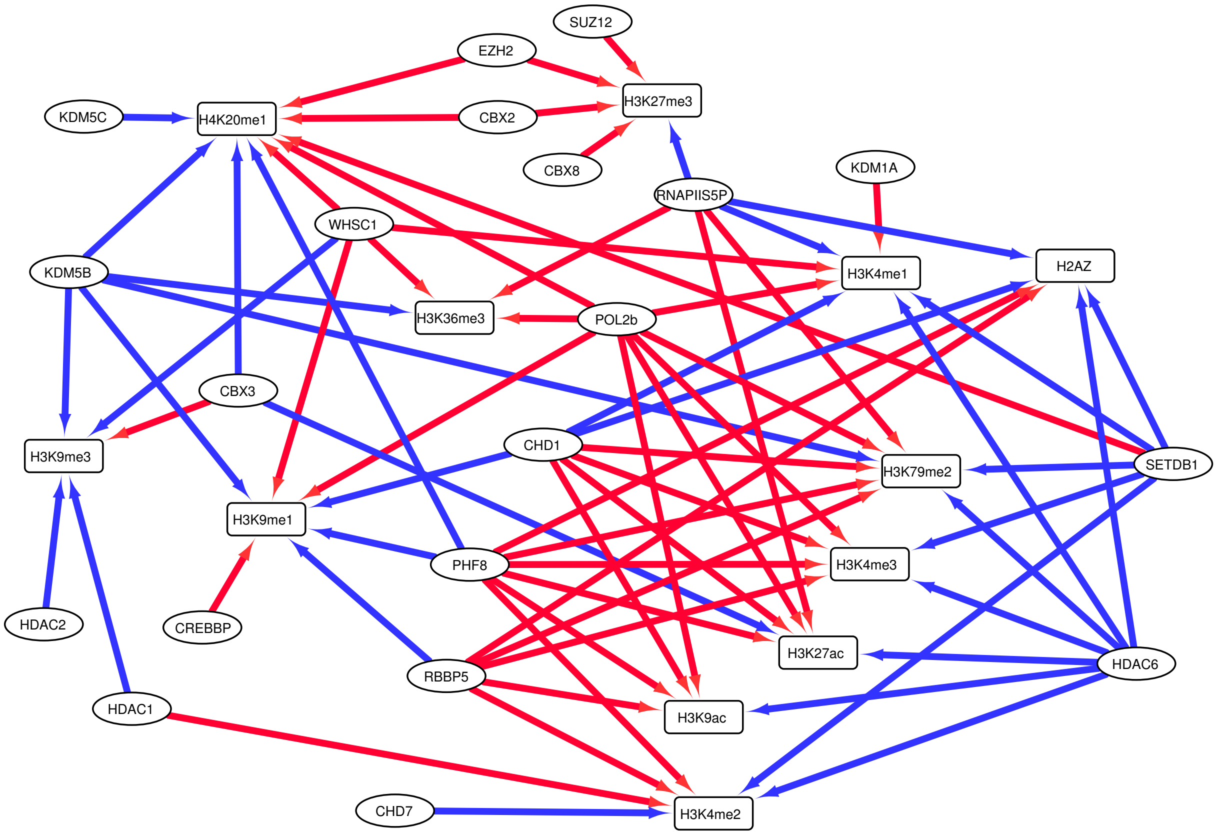

Figure 6.6.: *Interaction network between histone modifications (circles) and chromatin modifiers (rectangles) based on Elastic Nets. Red edges indicate positive interactions and blue edges indicate mutually exclusive interactions. The arrows do not indicate causation but rather indicate the direction of the prediction in the Elastic nets.*

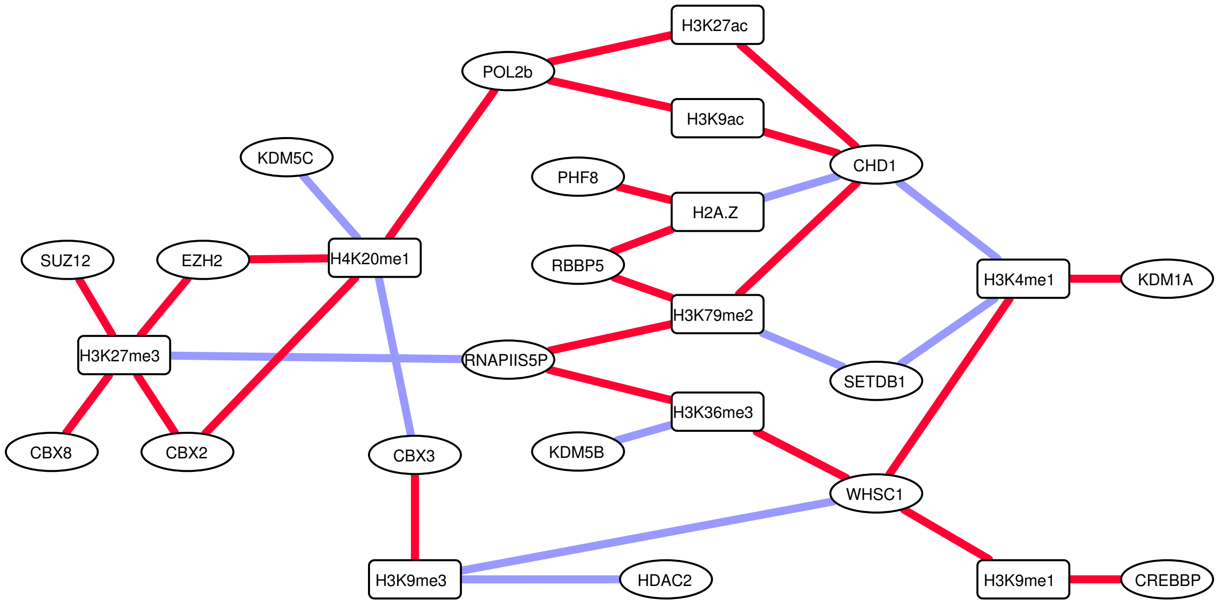

Figure 6.7.: *Consensus network of histone modifications (circles) and chromatin modifiers (rectangles) based on Elastic Nets and SPCN. Red edges indicate positive interactions and blue edges indicate mutually exclusive interactions.*

scribed in section 6.3.2 leading to a very sparse network. Many interactions (19 out of 33) are supported by literature evidence where the link between an HM and a CM is validated biochemically or via one protein that is not observed in our data set (Table 6.1). Our interactions complement the experimental evidence that has been obtained either *in vitro* or by using one or few genes as model system. Together, this large overlap with existing data suggests that also the unknown interactions are biologically relevant and could be the basis for new biological hypotheses. Indeed, based on our analysis the positive interaction between H4K20me1 and the Polycomb repressive complex has been validated experimentally[2] using a Co-IP experiment [1].

6.4. Discussion of novel interactions in the light of existing literature

From the existing literature and the novel interactions predicted by our network we can derive biological hypothesis about the function of HMs and CMs. Each edge has three possible mechanistic interpretations: In case of a positive association between a CM and an HM, the CM may be involved in (1) depositing the HM, (2) maintaining or interfering with the removal of the HM, or (3) it may be recruited by the HM. Similarly, for a negative association, the CM may be involved in (1) removing the HM, (2) counteracting the deposition of the HM, or (3) it may be repelled by the HM. Here, we will investigate three novel interactions in the light of existing literature.
CHD1 is positively connected to H3K9ac and H3K27ac in the network indicating that CHD1 tightly associates with H3K9ac/K27ac and that this association is independent of the histone acetyltransferases (HATs) responsible for the acetylation. These HATs

[2] Experiments performed by Dr. Sarah Kinkley

(KAT2A and EP300, respectively [94]) are also part of the set of CMs used in our analysis, but we do not detect an significant interaction. In line with this observations, the HATs are released from their targets after autoacetlylation (reviewed in [95]). CHD1 is recruited to chromatin by binding to H3K4me3 [96, 97]. A direct influence of H3K9ac/K27ac on the binding of CHD1 to chromatin seems unlikely because e.g. H3K9ac has no effect on the binding of CHD1 to H3K4me3 [97]. Thus, a possible hypothesis might be that CHD1 is involved in maintaining H3K9ac and H3K27ac at active promoters. That in turn could be part of a mechanism to prevent the H3K9 methylation of these residues and hence heterochromatin formation. This hypothesis is in line with several publication [96, 98].

In the network we find a positive association between RBBP5 or PHF8 and H2A.Z levels and a negative association of CHD1 and H2A.Z. The knockdown of H2A.Z in mESCs leads to reduced levels of RBBP5 concomitant with reduced levels of H3K4me3. Further, reduced levels of H3K4me3 lead to a decrease of RBBP5 as well as H2A.Z levels [73]. An appealing hypothesis would thus be that H2A.Z, RBBP5 and H3K4me3 form a positive feedback loop, where H2A.Z acts upstream of RBBP5, RBBP5 upstream of H3K4me3 and H3K4me3 signals back to H2A.Z. In contrast, PHF8 can bind H3K4me3 [99–102] and acts as a demethylase for H4K20me1, H3K9me1/2 and H3K27me2 [103]. An appealing hypothesis might be that PHF8 acts downstream of H3K4me3 to either promote H2A.Z deposition or repress H2A.Z removal. In line, PHF8 could indirectly promote H2A.Z position through demethylating residues that are then available for acetylation. Acetylation, in turn, is required for H2A.Z deposition [104]

We find a positive association between H4K20me1 and the Polycomb members CBX2 and EZH2. H4K20me1 is known to be tightly regulated during cell cycle with highest levels during mitosis [105]. A possible hypothesis would be that H4K20me1 and H3K27me3 guide the Polycomb repressive complex to their target sites throughout replication, which poses major challenges to the maintenance of Polycomb-mediated repression. In line with this hypothesis, H4K20me1-binding proteins that are able to interact with Polycomb members (e.g. [106]), but for which no ChIP-seq data was available, could mediate this interaction.

6.5. Summary

In summary, we have analyzed the most comprehensive set of genome-wide ChIP-Seq data in human cells to explore the chromatin signaling at different levels of granularity. First, we connected the set of available HMs and CMs to gene expression. Second, we have detected modules of HMs and CMs connecting sets of CMs to sets of functionally related HMs. Third, we have resolved these modules to highlight pathways connecting different regulatory states of the chromatin. Finally, we have applied an approach based on Elastic Nets and SPCN to identify specific co-localization pattern between HMs and CMs. This approach identified many known but also unknown interactions. Taken together, starting from co-localization patterns and applying methods that recover only high-confidence interactions we are able to move closer towards a mechanistic understanding of the interactions between HMs and CMs.

Table 6.1.: *A summary of the evidence in the literature for the links present in the consensus network (Figure 6.7). The literature evidence (column "Reference") can be indirect via maximal one other protein for which ChIP-Seq data was not available.*

HM	**CM**	**Reference**	**Comment**
H2A.Z	CHD1	[72]	CHD1 is involved in the removal of H2A.Z in the gene body
H2A.Z	PHF8	-	-
H2A.Z	RBBP5	[73]	H2A.Z recruits RBBP5
H3K27ac	CHD1	-	-
H3K27ac	Pol2b	[74, 75]	Elongator complex carries HAT activity, Indicator of active promoter
H3K27me3	CBX2	[76]	CBX2 binds to H3K27me3
H3K27me3	CBX8	[76]	CBX8 binds to H3K27me3
H3K27me3	EZH2	[76–79]	EZH2 catalyzes H3K27me3
H3K27me3	SUZ12	[80, 81]	SUZ12 binds H3K27me3 via EED
H3K27me3	RNAPIIS5P	[82]	High levels of RNAPIIS5P indicate elongation, which is repressed by H3K27me3
H3K36me3	KDM5B	-	-
H3K36me3	RNAPIIS5P	[83]	Pol II carries SETD2 which methylates H3K36
H3K36me3	WHSC1	[84, 85]	WHSC1 methylates H3K36
H3K4me1	CHD1	-	-
H3K4me1	KDM1A	[86, 87]	KDM1A demethylates H3K4me1
H3K4me1	SETDB1	[88]	SETDB1 binds less well to H3K4me1 substrates
H3K4me1	WHSC1	[89]	WHSC1 binds to H3K4 unmodified
H3K79me2	CHD1	-	-
H3K79me2	RBBP5	[90, 91]	May reflect the effect of RBBP5-mediated H2BK120ub, that boosts DOT1L activity
H3K79me2	RNAPIIS5P	[90]	Pol II carries DOT1L which methylates H3K79
H3K79me2	SETDB1	-	-
H3K9ac	CHD1	-	-
H3K9ac	Pol2b	[74, 75]	Elongator complex carries HAT activity, Indicator of active promoter
H3K9me1	CREBBP	-	-
H3K9me1	WHSC1	-	-
H3K9me3	CBX3	[14, 92]	CBX3 binds to H3K9me3
H3K9me3	HDAC2	-	-
H3K9me3	WHSC1	[89]	WHSC1 binds less well to H3K9me3 substrates
H4K20me1	CBX2	-	-
H4K20me1	CBX3	-	-
H4K20me1	EZH2	-	-
H4K20me1	KDM5C	-	-
H4K20me1	Pol2b	[93]	Pol II carries SETD8 which methylates H4K20

CHAPTER 7

Comparing regulatory mechanisms across chromatin states

The cell-type specific regulatory state of the genome can be observed by investigating the local chromatin environment, also called *chromatin state*. Specific combinations of histone modifications and cytosine modifications specify different chromatin states (see Section 2.2). However, the functional impact of the different combinations is not yet clear. Especially their implications on the recruitment of chromatin-related proteins to a specific chromatin environment are mostly unknown.

In the following chapter, we integrate 139 experiments including ChIP-Seq and MeDIP experiments from various sources to investigate the interactions between chromatin modifications and chromatin-binding proteins. In contrast to Section 6, where we investigated only the promoter region, in the following section we focus on comparing the regulatory mechanisms across chromatin states. We aim at identifying relevant interactions that characterize a particular chromatin state and investigate commonalities in the different chromatin states to infer possible signaling pathways that manifest a certain regulatory and chromatin state.

7.1. Data collection

A plethora of genome-wide data sets from different laboratories is available in mouse embryonic stem cells (mESC). These data sets comprise various ChIP-Seq data on HMs and chromatin-binding proteins, as well as MeDIP data on three different cytosine modifications, i.e. 5-hydroxymethylcytosine (5hmC), 5-methylcytosine (5mC) and 5-formylcytosine (5fC). We collected a set of relevant experiments from the GEO database (summarized in Table B.2). Our collaborators processed the experiments uniformly to avoid technical biases from the computational pre-processing pipeline. In detail, the reads are aligned to the mm9/NCBI37 reference genome with BWA allowing at maximum one mismatches and retaining only uniquely mapping reads. The resulting set of samples comprise 13 HMs, 3 cytosine modifications and 61 chromatin-binding proteins (see Ta-

ble B.2).
Since we are mixing samples produced in different labs, we check the genome-wide consistency of the available technical replicates and of proteins with known similar functions. We segment the whole genome into non-overlapping 200*bp* bins and select all bins that are classified as enriched after the binarization using the ChromHMM procedure (see Section 7.2). We apply hierarchical clustering using the Pearson correlation coefficient as a distance measure to the matrix consisting of read counts within the selected 200*bp* bins across all available samples. Figure A.15 shows that the technical replicates mostly cluster together or are at least in the same branch of the tree. Further, proteins with known common function, e.g. the Polycomb repressive complex, cluster together in the same branch. This suggests that the technical replicates and the different experiments are consistent on a genome-wide scale.

7.2. Defining chromatin states

Since the chromatin state largely depends on the particular cell type and cellular state, there is no gold-standard chromatin state annotation available. Functional classification of genomic elements into various chromatin states is usually done by investigating combinatorial patterns of chromatin modifications. The ChromHMM software [107] learns and detects different chromatin states from ChIP-seq experiments of HMs. The method is based on a multivariate Hidden Markov Model (HMM) that describes a number of chromatin states by the frequency of observing different combinations of chromatin modifications together and the frequency of transitions between chromatin states along the genome. Using this HMM, the genome can be segmented into the previously defined chromatin states.
Our collaborators applied the ChromHMM software to the set of 13 HMs, 3 cytosine modifications and CTCF in mESCs. The cytosine modifications were measured with MeDIP experiments that are, similarly to ChIP-seq, also antibody-based and thus can be readily integrated into the ChromHMM. For the learning, we integrate a core set of chromatin features that define the chromatin itself. We do not include other proteins that interact with this core set and thus would only indirectly define the chromatin state. In detail, our collaborators segmented the genome into non-overlapping 200*bp*-bins. Each bin is then binarized independently based on the number of reads mapping to the bin using the standard procedure implemented in the ChromHMM software and a threshold of 10^{-4}. In the remainder of this section, we will refer to the bins that were called enriched by this procedure as peaks. The training of the HMM model is done by using between 20 and 33 randomly initialized chromatin states. Following previously established selection strategies [108], we settle on a model with 20 different chromatin states.
Using the resulting HMM, our collaborators compute the probability of each 200*bp*-bin to belong to one of the chromatin states. For each segment, we then assign the most probable chromatin state. For the following analysis, we require that each segment should contain at least one peak of any sample (including HMs, cytosine modifications or CMs) and that the posterior probability of the state should be larger than 0.95. This procedure reduces the number of genomic regions to 840.793 segments, which consist of consecutive bins of the same chromatin state, of varying size (with about 62% of the segments having length 200*bp*).

Table 7.1.: *Summary of known annotations or other cell-type specific data by chromatin state (column "Observations"). The column "Classifications" associates the different chromatin states to transcriptional processes, states or regulatory elements. The association is derived manually from literature knowledge about the combination of enrichments in column "Observations".*

State	Classification	Observations
1	initiation	H3K79me2; gene body; RNA-Seq
2	elongation	H3K36me3; 5hmC; 5fC; exons; RNA-Seq
3		
4		H3K36me3; gene body; RNA-Seq
5		
6	heterochromatin	H3K9me3; H3K20me3; LaminB1
7		repeats; LaminB1
8		5fC; H3K20me3; H3K36m2; repeats
9		empty
10		5mC; 5hmC; repeats; LaminB1
11	inactive gene	5mC; 5hmC; H2AZ; gene body; DNase-tags
12	inactive promoter or enhancer	5hmC; H3K4me1/2; H2AZ; TSS; DNase-tags; ChIA-Pet interactions; CAGE-tags
13	enhancers	H3K4me1; intergenic; ChIA-Pet interactions
14		H3K4me1; H3K27ac; intergenic; DNase-tags; ChIA-Pet interactions; CAGE-tags
15	active promoter	H3K4me1/2; H3K27ac; H3K79me2; upstream of (active) TSS; RNA-Seq
16		H3K4me3; H3K27/K9ac; (active) TSSs; DNase-tags; ChIA-pet interactions; CAGE-tags; CpG-dense regions
17		H2AZ; H3K4me2; H3K27ac; precedes (active) TSS; DNase-tags; ChIA-pet interactions; CAGE-tags
18	Polycomb-repression	H3K27me3; H2Aub1; (repressed) TSSs; ChIA-pet interactions; CAGE-tags; CpG-dense regions
19		H3K27me3; precedes (inactive) TSS
20	CTCF	CTCF

7.3. Characterization of the chromatin states

The ChromHMM method is unsupervised making the functional implications of the detected states not directly apparent. Additionally, the core set of features that define our chromatin states is different from those used in previous applications. This could potentially lead to different chromatin state definitions compared to previous, well characterized ones. In order to reveal possible functional differences between the chromatin states, we compare the ChromHMM states to known annotation or other cell-type specific data in the following section. For convenience we refer to the 20 different states as indicated in the column "Classification" in Table 7.1. These names are based on the evidence discussed in more detail in the following section.

7.3.1. Chromatin modifications

The enrichment for the core features among each state is shown in Figure 7.1(A). It is calculated as the number of ChromHMM peaks of a certain feature in a particular state divided by the number of all segments in that state. Since we will later work with the normalized read counts instead of the peaks where information might be lost by the peak calling, we also investigate the enrichments of the normalized read counts in Figure 7.1(B). We scale the genome-wide distribution of the read counts to have mean zero and calculate the mean signal value in each chromatin state. The enrichments are mostly consistent between the ChromHMM peaks and the normalized read counts, but, as expected, also show slight differences, e.g. in the H3K27ac or the 5hmC. In the following, we interpret the states based on the consensus between the chromHMM and the read level enrichments. State 16 and, to a lesser extent, state 15 show enrichment in H3K4me3, H3K79me2, H3K27ac and H3K9ac, a characteristic of active promoters and initialization. In contrast, state 17 shows only enrichment for H3K4me3 and H3K4me2 and, similarly to state 16, enrichment for the histone variant H2AZ. State 18 and 19 are mainly enriched for H3K27me3 which is characteristic for Polycomb-mediated repression. However, state 18 is also enriched for H3K4me3 and for H2Aub1. States 12, 13 and 14 show enrichment mostly for H3K4me1, a mark related to enhancers. Additionally, state 14 is enriched for H3K27ac indicating that this state might mark active enhancers. State 1-5 are enriched for H3K36me3, a sign for elongation. The cytosine modification 5fC is most prominent in state 8, which is among other also characterized by H4K20me3 and H3K36me2. In contrast, 5hmC is distributed among several states and mainly present in state 2, 11 and 12. The classical 5mC is most prominent in state 10 but also in state 11.

7.3.2. Known annotations

Since the chromatin states are solely based on ChIP-Seq data, we use known annotations to characterize the states in more detail. We plotted the enrichment of the different chromatin states along the RefSeq genes (Figure 7.2) using the *EpiCSeg*-package in R. The genes are aligned at their TSS and the gene body is scaled to correct for different gene lengths. For each position we then count how many genes have a certain chromatin state at the corresponding position in the gene. Figure 7.2 thus summarizes the frequency of a certain state along the gene body and the surrounding regions. We observe clearly different patterns of enrichments for the different states. States 12, 16 and 18 are enriched directly at the transcription start site (TSS). Additionally, state 16 and 18 are enriched with CpG Islands (see Figure A.16) indicating that these states mark promoter regions. State 17 and state 19 are enriched upstream of the TSS, i.e. in the promoter proximal region. State 1 and 15 are enriched directly downstream of the TSS. State 2-5 are all enriched in the gene body. They only differ by their annotation with exons that are mostly enriched in state 2 and 3 (see Figure A.16). State 6, 9, 13, 14 and 20 are enriched outside of the gene body. State 7, 8, 10 and 11 are slightly enriched in the gene body but state 7, 8 and 10 also show enrichment for repeat regions (see Figure A.16).

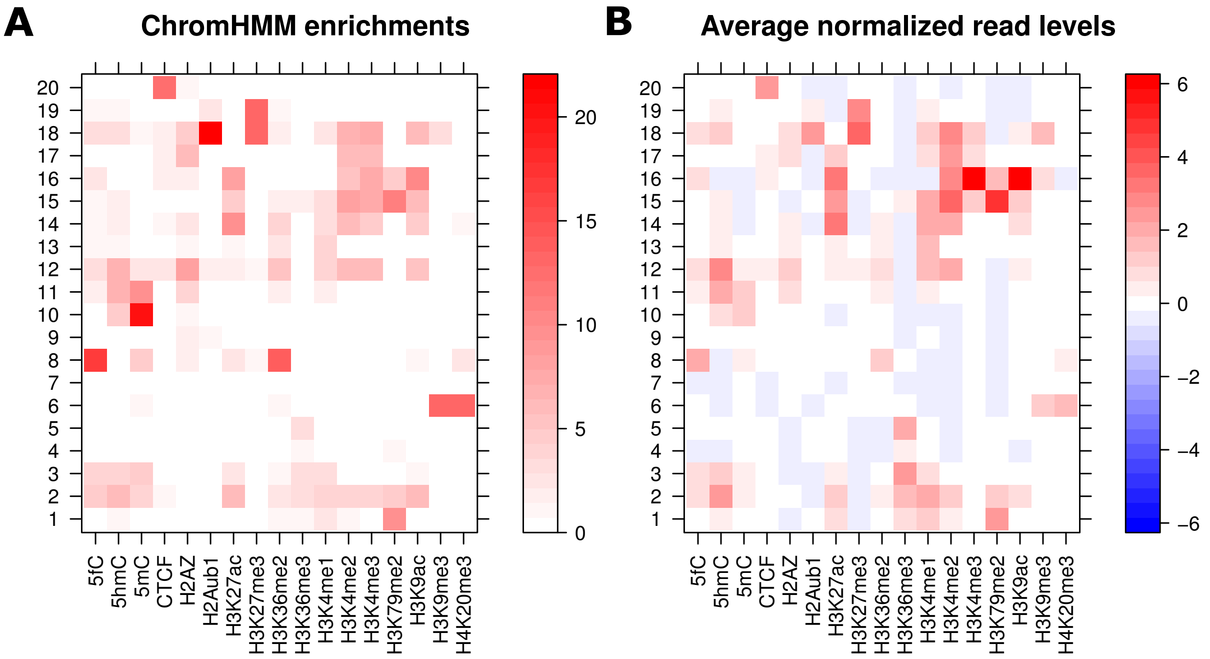

Figure 7.1.: *Heatmaps of the enrichments of the core features used to train the chromatin states. The enrichments were calculated based on (A) the chromHMM peaks and (B) the average normalized read signals in the chromatin state segments, respectively, for each chromatin state.*

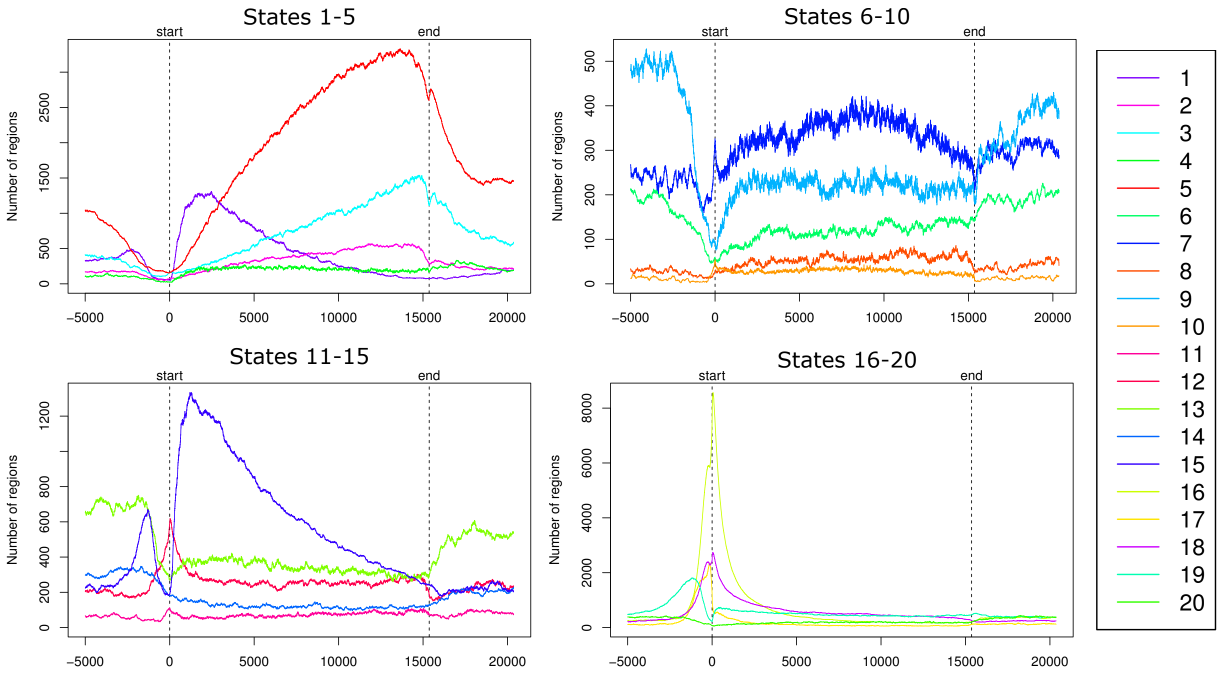

Figure 7.2.: *Distribution of the chromatin states across RefSeq genes. Genes are aligned at the TSS (start). The gene body is scaled proportionally to the gene length, s.t. the TES (end) of all genes are aligned too. For each position on the x-axis the number of bins having a particular state over all genes is plotted.*

7.3.3. Other cell-type specific data

Other data that describe the cell-type specific state of the genome can be used to further characterize the chromatin states. For example, state 1-5, as well as state 15 and 16 show signs of expression measured by RNA-seq (Figure A.16). DNase-hypersensitive regions, which characterize enhancer and promoter regions, are prominent in states 2, 12 and 14-18. LaminB1, which can be found at genomic elements that are associated with the nuclear lamina, is enriched in the states 6, 7 and 10 (Figure A.16). We observe PolII-mediated ChIA-Pet interactions, which also characterize enhancer and promoter regions, mostly in the states 16, 17 and 14 and slight enrichment in state 18, 12, 15 and 13. Finally, we investigate the enrichment for promoter and enhancer-specific CAGE-tags (see Figure A.16). We see a clear enrichment of the enhancer-specific CAGE-tags in state 14 and weaker enrichment in state 2, 12 and 15-18. In contrast, the promoter-specific CAGE-tags are highly enriched in state 16 and weakly enriched in state 17-18.

7.3.4. Transitions between states

Next, we investigate how the different states are spatially related to each other, i.e. whether certain states appear consecutively or independently of each other. We calculated the transition probabilities for all chromatin state segments. We make the transition matrix symmetric by counting the predecessor and the successor for each segment. Due to the filtering applied before (see section 7.2), the segments are not necessarily directly adjacent to each other. For a transition we allow a maximum distance of $1200bp$ between the start and the end of the consecutive segments. By this definition, on average 33% of the bins per state lack adjacent segments. The states 10, 20, 8, 7, 11 and 9 are reduced by at least 50%, while the promoter or Polycomb-repressed states 15, 18, 16, 1 and 19 retain most segments.

The matrix of transition probabilities is shown in Table B.3. We observe many off-diagonal elements that show transitions between states. Among them a frequent transition between the two initiation-related states 1 and 15, but while the more promoter-like state 15 also frequently transitions to the active TSS state 16, state 1 does not. State 16 itself shows frequent transitions to state 17, a promoter-like state. Together this suggest that there is a pattern of state 17 following state 16, following state 15 to 1, showing a sequential transition from an active promoter to initiation. Similarly, we find frequent transitions from the initiation state 1, to the elongation states 2 to 3 to 5. The heterochromatic states 8 and state 11 transition most frequently to the empty state 9 and to a lesser extent to the CTCF state 20. State 9 also transitions to state 13, marking inactive enhancers. Interestingly, many other states are enriched in transitions to the empty state 9. These include state 8, 11, 12, 13 and 20, which are all related to heterochromatic or inactive genomic elements. Among them state 11, which is related to inactive genes, frequently transitions to the inactive promoter or enhancer state 12 and state 12 transitions to inactive enhancer state 13. Additionally the active enhancer state 14 transitions frequently to the inactive enhancer state 13. In summary, this suggest patterns of transitions between states that resemble related transcriptional and regulatory programs.

7.3.5. Summary

We have identified 20 chromatin states from 13 HMs, 3 cytosine modifications and CTCF. By comparing them to known annotations and other cell-type specific chromatin-related data, we can assign potential functions to the segments assigned with different states. Since state 16 is most enriched in H3K4me3 and H3K27/K9ac, mostly found at (transcribed) TSS, shows many ChIA-pet interactions and contains CpG dense regions, we conclude that state 16 describes active TSS. State 15 is directly upstream of state 16, characterized by enrichment for H3K79me2 and located in the region directly upstream of the TSS. It might thus describe the first exon of actively transcribed genes. State 17 precedes state 16 and does not show acetylation. It thus might mark the promoter proximal region of actively transcribed genes. In contrast, state 19 is also located directly upstream of the TSS, but is mutual exclusive with state 17, shows signs of H3K27me3 and is transcriptionally silent. State 19 might thus characterize the proximal promoter regions of repressed genes. Further, segments of state 19 frequently follow segments of state 18, a state that is characterized by H3K27me3, most frequently occurs at the repressed TSS and contains many CpG islands. However, since we find also enrichment for H3K4me3, ChIA-pet interactions and CAGE-tags in state 18, this state might also contain bivalent promoters. Finally, state 13 and 14 are frequently observed outside of gene bodies and are marked with H3K4me1 and ChIA-Pet interactions. Additionally, state 14 shows enhancer associated CAGE-tags, as well as enrichment for H3K27ac, which together suggest that state 14 marks active enhancers.

7.4. Chromatin state-specific signaling networks

So far we investigated only the core chromatin features including histone and cytosine modifications to identify different chromatin states. These chromatin states are characterized by distinct combinations of the chromatin features and show different characteristics for known genome annotation, transcriptional profiles or cell-type specific chromatin features. However, subgroups of states seem to have commonalities in their chromatin regulatory program, as seen, for example, by shared enrichment for a particular chromatin modification. To infer functionally relevant contributions of the chromatin modifications in each of the states, we reconstruct chromatin-state specific interaction networks. As motivated in Section 6, these networks reveal the most relevant interactions between the chromatin modifications and the chromatin modifiers. Each network then connects the different combinations of chromatin features to specific chromatin-related proteins allowing to attribute potential chromatin state-specific functional roles to each component.

7.4.1. Data pre-processing

For the network reconstruction we count the reads falling within each chromatin segment. Each segment is normalized by its length. If technical replicates exist, they are summed up for both replicates of samples and input. The samples are normalized as described before (see Chapter 6) based on all segments independent of the chromatin state. For each state, we standardize the read counts to have a mean of zero and a standard deviation of one.

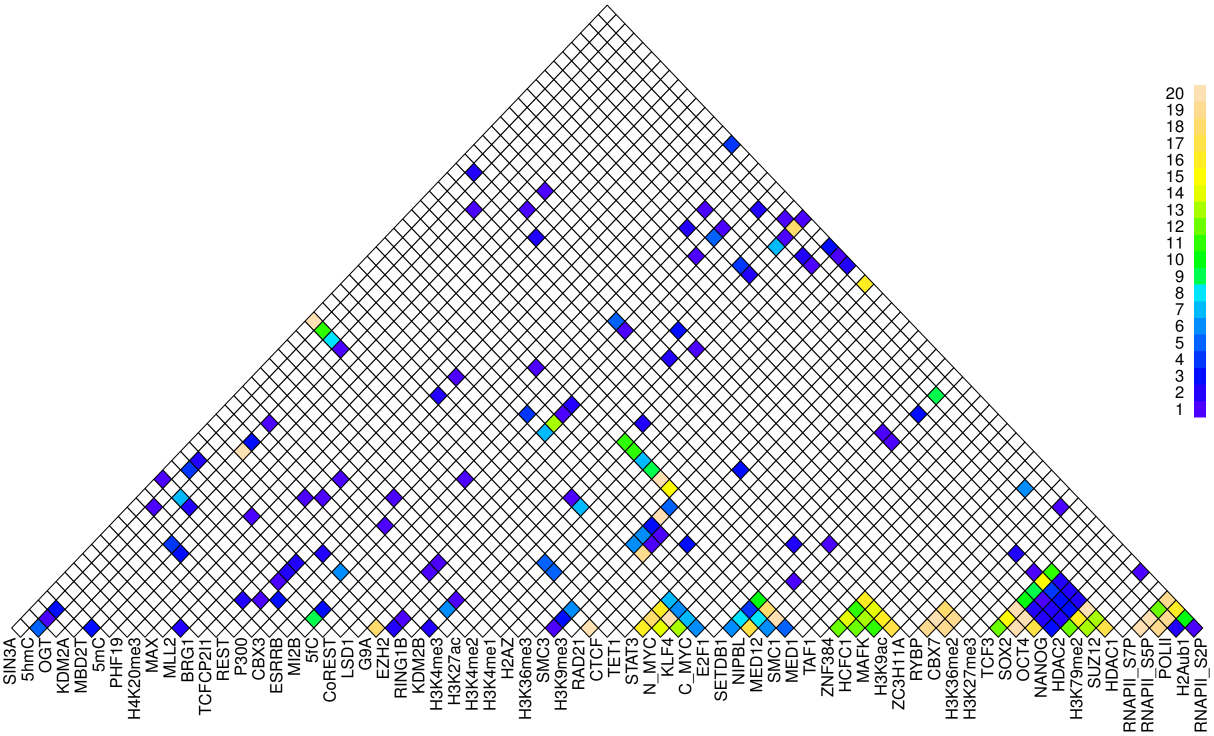

Figure 7.3.: *The number of states in which each edge is found. The color of each cell indicates the number of states the chromatin components indicated on the diagonal co-localize. White cells indicate that the corresponding edge is never predicted.*

7.4.2. Chromatin state-dependent network reconstruction

For each chromatin state, we reconstruct an interaction network with the same procedure as explained in Chapter 6. We construct the SPCN on all available samples based on the segments in each single state. For the Elastic Nets, we predict each HM, cytosine modification or CM by the remaining CMs. We return only those edges where the median of the coefficients in a 10-fold CV is larger than twice the standard deviation of the coefficients for each sample. The final network is constructed by combining the SPCN and the Elastic Nets for each state, leading to 20 different networks (summarized in Figure 7.3). The networks consist of 62 to 95 edges and connect between 52 and 62 nodes. In total there are 237 unique edges of which 193 edges reflect positive and 44 reflect mutual exclusive interactions.

Figure 7.3 shows the frequency of the interactions in the different chromatin states. While 31% of the interactions are specific for a certain state, there are about 5.9% of the interactions in common to all states and about 28.5% of the interactions shared by half or more of the states. Such interactions are, for example, formed by the transcription factors STAT3, MYC and KLF4 or TCF3, SOX2, OCT4 and NANOG. Further, we find that members of the mediator complex (MED1 and MED12) frequently co-localize with NIPBL and members of the cohesin complex (SMC1 and SMC3). Also the various forms of PolII phosphorylation frequently co-locate with each other, but only PolII-S2P frequently interacts with H3K36me3. TET1 forms interactions with SIN3A and 5hmC independently of the states. TAF1 co-occurs frequently with histone marks known to be present at active promoters. The interactions between RYPB, CBX7 and H3K36me2 or H3K27me3 also seem to be independent of the states.

7.4.3. Comparing the chromatin state-specific signaling networks

The union of all chromatin state-dependent interactions identified in the previous section results in a chromatin signaling network summarizing all specific interactions of each chromatin modification and chromatin modifier. Due to the chromatin state-specific reconstruction, we expect that different parts of the network are used at different chromatin states. In the following section, we focus on identifying commonalities in the usage of the chromatin signaling network across chromatin states. This leads to a systems-level view on the chromatin signaling network relating, for example, shared parts of the network to similar chromatin states. Note that, instead of looking at all possible interactions, we focus only on the interactions that were most specific to a certain chromatin modification or modifier.

To quantify which chromatin-signaling interaction is used in which state, we compute for each interaction the enrichment as the fraction of $200bp$ bins in a chromatin state showing a peak for both interaction partners. Thus, for each chromatin state we obtain an enrichment profile showing the enrichment of all interactions in this state. Figure 7.4 shows the matrix of the Pearson correlation coefficients between the enrichment profiles ordered by hierarchical clustering. By cutting the tree resulting from the hierarchical clustering as indicated in Figure 7.4, we see 7 different groups of states that show highly correlated enrichment profiles of the chromatin signaling interactions. As expected, the expression-related states 1-5 show similar enrichment patterns. Further, the Polycomb-related states 18-19 are very correlated. However, while state 18 is also correlated with state 12 and 16, state 19 is anti-correlated with state 16. Interestingly, the active enhancer state 14 is grouped together with the active promoter states 15-17, while the inactive enhancer state 13 clusters together with the states 6, 7 and 9. Another group of states is formed by states 10-12.

Interestingly, state 8, which is characterized by strong enrichment for 5fC, is very different from all other states and similarly to state 19 shows anti-correlation with the active promoter states. In contrast, state 12, which is highly enriched with 5hmC, shows similar interaction enrichment patterns as the Polycomb-states and the active promoter states. The 5mC is mostly present in states 10 and 11 that show similar interactions as state 8 and 12 but neither correlate nor anti-correlate with the active promoter states.

7.4.4. Characteristic sub-networks define groups of chromatin states

In the previous section, we identified groups of related chromatin states based on the enrichment profiles calculated from the co-occurrence of the two interaction partners that participate in a chromatin-signaling interaction. In the following section, we investigate which interactions contribute to the observed similarities within the groups of related states. For each interaction, our collaborators calculate an enrichment profile within the chromatin states as described in Section 7.4.3. The enrichment profile was scaled to have a mean of zero and standard deviation of one. The resulting matrix was used to perform hierarchical clustering on the interactions. Using this procedure we detect distinct groups of interactions, called *sub-networks* in the remainder, that show similar enrichment profiles over the chromatin states. Note, that these sub-networks are more detailed than clusters of proteins that show high pair-wise correlation, since we pre-selected the interactions based on the global chromatin signaling network.

Our collaborators manually annotated 15 sub-networks (see Figure A.17) that contain

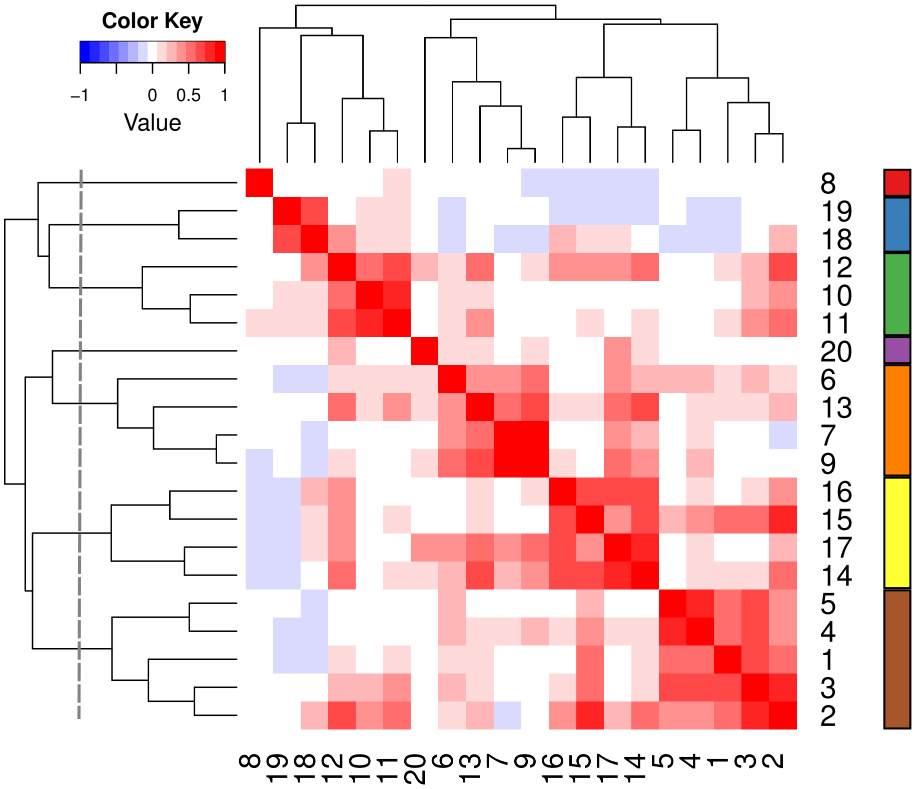

Figure 7.4.: *Heatmap of the Pearson correlation between the states based on the enrichment of all interactions across the genome. The enrichment of each interaction was calculated as the fraction of* 200*bp bins in a chromatin state showing a peak for both interaction partners. We derive groups of similar states by cutting the tree resulting from the hierarchical clustering of the correlation profiles at the dashed line. The colored bars on the site indicate the resulting groups of states.*

between 2 and 30 interactions and connect between 4 and 27 nodes. Some sub-networks show interactions between known complex members. For example, sub-network 8 shows interactions between Polycomb complex members and repressive histone marks. Sub-network 3 shows interactions between the members of CTCF/cohesin and the Mediator complex. Also sub-network 4 shows interactions between members of the mediator complex, as well as the NuRD complex. Other sub-networks reflect interactions of known chromatin states. For example, sub-network 2 and 11, as well as sub-network 15, show signs of active promoters and initiation. Sub-network 5 shows elongation. Finally, sub-network 7 connects the pluripotency transcription factors to the Co-/REST complex. Interestingly, the two sub-networks 13 and 14 comprise interactions of which most involve 5fC and 5hmC, respectively.

The sub-networks are enriched in different chromatin states (see Figure 7.5). Some sub-networks show strong enrichment in the groups of chromatin states identified in Figure 7.4. The Polycomb-related sub-network is mostly present at state 18 and 19. The group of chromatin states 10-12 is characterized by sub-network 14. The two sub-networks 4 and 7 characterize the group of states including state 6,7,9 and 13. The elongation-related sub-network 5 is enriched at all states marking transcribed gene bodies. Finally, the states involving active promoters and enhancers (states 14-17) are mainly characterized by sub-network 4 and 15. Thus the main similarities between the states can be captured

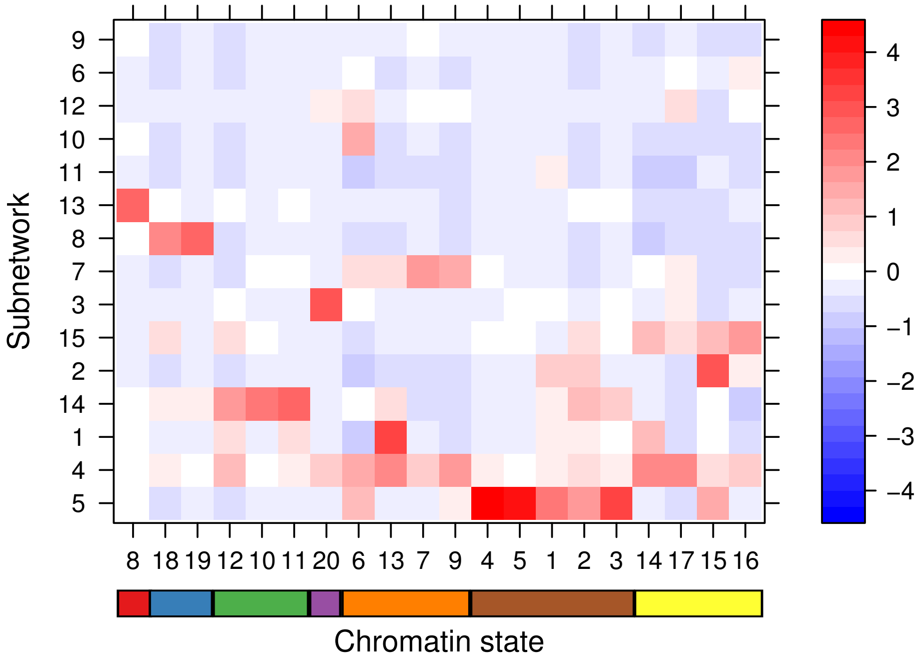

Figure 7.5.: *Heatmap of the enrichments of the sub-networks across all chromatin states. The enrichment indicates the average enrichment of the interactions in each sub-network. For each interaction the enrichment across the chromatin states was z-transformed. Sub-networks were reordered by hierarchical clustering on the average enrichment profiles. Chromatin states were reordered to match the order in Figure 7.4 and the colored bars indicate the groups of states as identified in Figure 7.4.*

by a few sub-networks.

Individual states also show state-dependent enrichment for a particular additional sub-network. For example, although the Polycomb-related states 18-19 are both characterized by the Polycomb sub-networks, state 18 shows also enrichment for the promoter-related, the 5hmC-related and mediator related sub-network 15, 14 and 4 respectively. In contrast, state 19 only shows enrichment for the 5hmC-related sub-network 14. Also the group of states including state 10-12, which are mostly characterized by the presence of sub-network 14, show differential enrichment for other sub-networks. While state 10 is only enriched for sub-network 14, the other two show also enrichment for the mediator-related and the H3K4me1-related sub-network 4 and 1. Additionally state 12 also shows enrichment for the promoter-related sub-network. This indicates that although some states share a main set of interactions between chromatin modification and chromatin-related proteins, the individual states have additional regulatory programs. The promoter associated states 15-17 show besides the common enrichment for the mediator sub-network 4 and the active transcription-related sub-network 15 also enrichment for other sub-networks. For state 15, these are the sub-networks 5 and 2, which are related to elongation. For state 16, the sub-network 6 related to MYC, E2F1 and KLF4, as well as the sub-network 2 related to elongation are enriched. Finally, state 17 is also characterized by an enrichment for sub-network 7 (pluripotency factors and NuRD-/REST complex members) and sub-network 12 (SETDB1).

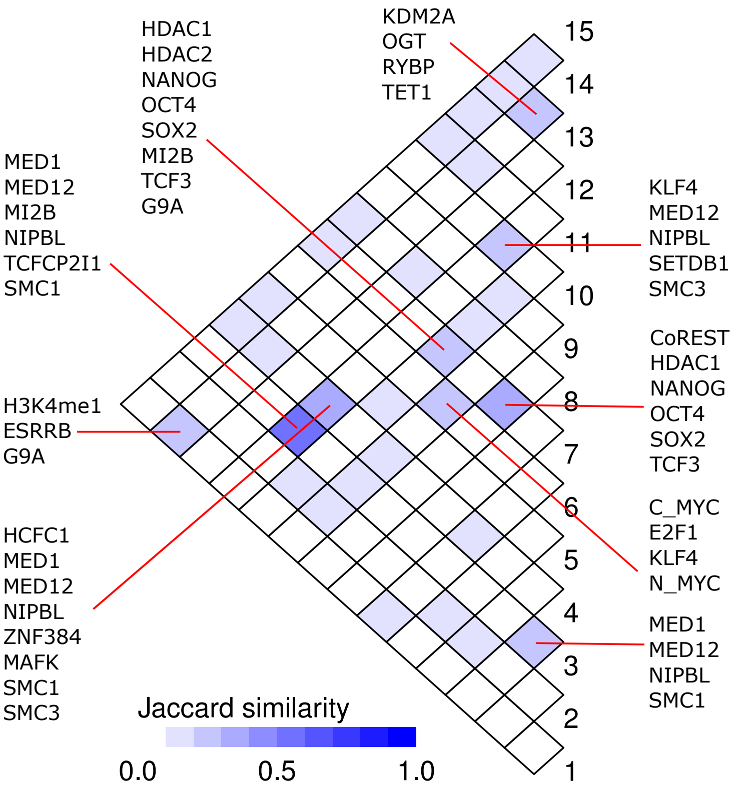

Figure 7.6.: *Heatmap of the overlap between the proteins in the sub-networks. The overlap is measured by the ratio between the cardinality of the intersection of proteins and the cardinality of the union of proteins within two sub-networks.*

7.5. Interactions within and between sub-networks

In section 7.4, we investigated similarities in the chromatin-related interactions among the different chromatin states. We have seen that certain groups of chromatin states share similar interaction patterns. Distinct sub-networks characterize these chromatin state groups. However, individual chromatin states also show enrichment for additional sub-networks. In the following section, we will characterize the different sub-networks in more detail to gain a deeper understanding of the unique regulatory state of each chromatin state.

7.5.1. Sub-networks share interaction partners

Although most sub-network contain distinct sets of interactions, many proteins are shared across the sub-network. Only sub-network 3 (CTCF and RAD21), 8 (PHF19), 10 (H3K9me3, H4K20me3 and STAT3), 11 (H2AZ and TAF1), 13 (5fC, MAX and H3K36me2) and 14 (5hmC and 5mC) contain unique proteins or chromatin modifications.

Figure 7.6 indicates the overlap in proteins calculated as the fraction between the number of proteins shared among the sub-networks and the number of proteins in the union of the sub-networks. We see that there are several sub-networks that share many proteins. For example, sub-network 3, 4 and 12 share complex members of the mediator complex and other scaffolding proteins. sub-network 9 and 7 share members of the REST complex and the pluripotency transcription factors. Also sub-network 11 and 7 share the pluripotency transcription factors but also members of the NuRD complex. The sub-network 14 and 13 share proteins related to repression and oxidation of cytosine modifications.

7.5.2. Context-specific interactions of shared interaction partners

Many sub-networks share proteins or chromatin modifications but by definition the interactions of these proteins within the sub-networks are different. Since there are sub-networks that are very chromatin-state specific, we investigated the distribution of the interactions of individual proteins across the sub-networks. We find that most proteins are central to a specific sub-network but show a few interaction also in other sub-networks (see Figure A.18). For example, ESRRB interacts most within sub-network 4 but has also an interaction in sub-network 1 and 14. Similarly, H3K4me1 has most interaction partners in sub-network 1 but also in sub-network 14, 11 and 8.

Further, although certain groups of proteins mostly interact within the same sub-network, their individual interactions differ between sub-networks. For example, while LSD1 (also known as KDM1A), NIPBL, MED1/12, SMC1 and ESRRB all have most of their interactions within sub-network 4, each of these proteins shows interactions in distinct other sub-networks. LSD1 also interact with proteins in sub-network 7, 14 and 15. The mediator complex related proteins interact within sub-network 12 and 3.

7.6. Discussion

In summary, we have integrated public genome-wide DNA-binding data and applied network reconstruction in order to detect chromatin state-specific interactions among histone modifications, cytosine modifications and chromatin modifiers. We identified 20 chromatin states defined by unique patterns of histone modifications and cytosine modifications that correlate with various genomic and regulatory annotations. We connect the different chromatin states to the relevant chromatin modifiers by reconstructing chromatin state-specific interaction networks. These networks summarize the most important interactions between the chromatin modifications and chromatin modifiers. Further, the networks show unique state-specific interactions but also share enrichment for other interactions. In particular, a few distinct sub-networks, which summarize specific interactions that have a similar enrichment profile over the different chromatin states, capture most of the similarity between sub-groups of chromatin states. Depending on the chromatin state, these sub-networks are complemented with other chromatin state-specific sub-networks adding an additional regulatory unit which could potentially lead to the observed differences in the chromatin states. The identified sub-networks comprise interactions among known complex members or interactions typical for a particular transcriptional state. Interestingly, individual sub-network members, which mainly interact within a certain sub-network, also show interactions in other sub-networks. An appealing hypothesis would be that such proteins act as communicators between the chromatin states. As such, these proteins might recruit relevant complexes or regulatory programs that launch a transition to another chromatin states. However, additional experiments and analysis is needed to verify this hypothesis.

CHAPTER 8

Conclusion and discussion

The present thesis dealt with reconstructing chromatin signaling interactions from genome-wide DNA-protein binding data. We integrated several publicly available data sets to infer specific co-localization patterns relevant for the regulation of transcription. With the help of the detected interactions we derived testable biological hypothesis about transcriptional regulation and the role of histone modifications, cytosine modifications and chromatin modifiers within this process. We approached three main research questions which we will summarize and discuss in the following chapter.

A recurrent question in network reconstruction is which method performs best for a specific research question. In chapter 5, we compared several network reconstruction approaches that have been applied in the context of reconstructing chromatin signaling from genome-wide DNA-binding data. The methods compared in the present thesis include pairwise association-based approaches, several (sparse) Gaussian graphical, a Bayesian network and Maximum entropy-based approaches. Our analysis on simulated and on biological data did not reveal any specific advantage of any of the methods under the settings tested. The main difference between the methods already becomes apparent on the conceptual level. For example, the pairwise association-based approaches do not take into account any other data that is available and thus lead to many high-scoring false positive interactions. In contrast, the sparse methods, especially SPCN, are very conservative in detecting edges.

Only recently more and more genome-wide DNA-protein binding data sets comprising chromatin modifiers in human cell lines became available. In Chapter 6, we used the first comprehensive data set on chromatin modifiers in human cells to investigate the interactions between HMs and CMs at promoters and their role in transcriptional regulation. We found that both the levels of HMs and CMs at the promoter region are equally predictive for gene expression, suggesting that both act in the same chromatin signaling pathway relevant for transcription. We investigated the chromatin signaling network connecting HMs with CMs at different levels of detail. First, we detected different, functionally related modules of HMs and CMs. Then, we found that individual CMs show interactions with distinct set of HMs, suggesting a role of individual CMs in defining and maintaining a particular chromatin state. Finally, we detected, besides many interactions known in the literature, also new ones that could form the basis for biological and mechanistic hypothesis on the role of particular HMs and CMs in chromatin signaling.

Individual research projects usually study a few genome-wide DNA-protein binding based on a specific hypothesis. This data is available in public repositories allowing an integrative analysis without being driven by a specific hypothesis. In chapter 7, we integrated a large set of genome-wide DNA-binding data on HMs, cytosine modifications, TFs and CMs in mESCs to investigate chromatin-state specific interactions among the different components. The different chromatin states usually have state-specific interactions among HMs, CMs and cytosine modifications, although groups of chromatin states show similar enrichment patterns for different interaction sub-networks. Only a few proteins connect the different sub-networks by an interaction, suggesting that these proteins might be involved in communicating a certain chromatin state transition to other complexes.

Our understanding of the chromatin signaling pathways and their connection to transcriptional regulation is still incomplete. With the present thesis, we addressed two different aspects, as explained in the following, to improve the understanding of chromatin signaling. First, we investigated the individual interactions by integrating various genome-wide DNA-binding data and applying computational means to get as close as possible to the underlying mechanistic interactions. This kind of analysis helps to derive novel biological hypothesis about the genome-wide interactions of individual chromatin-related proteins. Only recently (e.g.[13]), these endeavors have been complemented by high-throughput proteomics approaches that identify interaction partners of specific chromatin modifications. However, with the data currently available these analyses cannot predict the cause or consequence of a binding event or whether it results from physical contact of the two chromatin components. The necessary global perturbation experiments to answer this kind of question are currently hampered due to redundancy in the regulatory mechanisms or lethality resulting from the perturbation. Nevertheless, *in vitro* or gene-based functional studies could be used to investigate the mechanistic implications of the derived interactions in more detail. Second, we analyzed large sets of publicly available data for general genome-wide trends that shed light on the broader function of HMs or CMs on the chromatin-signaling. This analysis shed light on the modular organization of the HMs and CMs that are combined to achieve a certain chromatin state. Further, we identified individual ”connector” proteins or modifications that seem to be involved in the context-specific recruitment of the modules. This systems-level view of chromatin signaling allows to understand the complex behavior of the chromatin signaling pathways. This could be especially useful for the understanding of complex diseases evolving from aberrant behavior of chromatin-related proteins.

In the present thesis, we focused only on interactions relevant to chromatin signaling. In future research these signaling processes have to be integrated into the whole cellular context. In particular, it would be interesting to investigate particular TFs that might recruit specific chromatin signaling modules and pathways to specific sites in the genome upon activation or in response to the cellular environments. Further, chromatin modifiers themselves are exposed to different other cellular signaling pathways leading, for example, to acetylation or phosphorylation of the protein. This ultimately would lead to enhanced or decreased activity of the chromatin modifier and thus to modifications in the cell-type specific, local chromatin states. Investigation of the response of the chromatin modifier to such cellular signaling pathways would shed further light on how the dynamic, cell-type specific state of the genome is regulated and manifested.

Bibliography

[1] J Perner et al. “Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling”. In: *Nucleic Acids Research* 42.22 (2014), pp. 13689–13695.

[2] J Perner and HR Chung. “Chromatin signaling and transcription initiation”. In: *Frontiers in Life Science* 7 (1 2013), pp. 22–30.

[3] E Carrillo de Santa Pau et al. *Functional analysis and co-evolutionary model of chromatin and DNA methylation networks in embryonic stem cells*. BioRxiv, 2014. DOI: `http://dx.doi.org/10.1101/008821`.

[4] O Ram et al. “Combinatorial Patterning of Chromatin Regulators Uncovered by Genome-wide Location Analysis in Human Cells”. In: *Cell* 147.7 (2011), pp. 1628 –1639.

[5] CD Allis et al. *Epigenetics*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 2007, pp. 29–32,40.

[6] Wikipedia Commons. *Chromatin Structure*. 2015. URL: `http://commons.wikimedia.org/` (visited on 04/01/2015).

[7] K Struhl. “Histone acetylation and transcriptional regulatorymechanisms”. In: *Genes & Development* 12.5 (1998), pp. 599–606.

[8] M Tahiliani et al. “Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1”. In: *Science* 324.5929 (2009), pp. 930–935.

[9] S Ito et al. “Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine”. In: *Science* 333.6047 (2011), pp. 1300–1303.

[10] M Weber and D Schbeler. “Genomic patterns of {DNA} methylation: targets and function of an epigenetic mark”. In: *Current Opinion in Cell Biology* 19.3 (2007), pp. 273 –280.

[11] WA Pastor et al. “Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells.” In: *Nature* 473.7347 (2011), 394397.

[12] EA Raiber et al. “Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase”. In: *Genome Biology* 13.8 (2012), R69.

[13] M Iurlaro et al. “A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation”. In: *Genome Biology* 14.10 (2013), R119.

[14] AJ Bannister et al. “Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain”. In: *Nature* 410.6824 (2001), pp. 120–124.

[15] Y Li et al. “Effects of tethering HP1 to euchromatic regions of the Drosophila genome”. In: *Development* 130.9 (2003), pp. 1817–1824.

[16] PJ Verschure et al. “In Vivo HP1 Targeting Causes Large-Scale Chromatin Condensation and Enhanced Histone Lysine Methylation”. In: *Molecular and Cellular Biology* 25.11 (2005), pp. 4552–4564.

[17] C Pouponnot, L Jayaraman, and J Massagu. “Physical and Functional Interaction of SMADs and p300/CBP”. In: *Journal of Biological Chemistry* 273.36 (1998), pp. 22865–22868.

[18] X Liu et al. “c-Myc Transformation Domain Recruits the Human STAGA Complex and Requires TRRAP and GCN5 Acetylase Activity for Transcription Activation”. In: *Journal of Biological Chemistry* 278.22 (2003), pp. 20405–20412.

[19] W Yuan et al. “Dense Chromatin Activates Polycomb Repressive Complex 2 to Regulate H3 Lysine 27 Methylation”. In: *Science* 337.6097 (2012), pp. 971–975.

[20] AM Khalil et al. “Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression”. In: *Proceedings of the National Academy of Sciences* 106.28 (2009), pp. 11667–11672.

[21] EM Mendenhall et al. “GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells”. In: *PLoS Genet* 6.12 (2010), e1001244.

[22] NP Blackledge et al. “CpG Islands Recruit a Histone H3 Lysine 36 Demethylase”. In: *Molecular Cell* 38.2 (2010), pp. 179 –190.

[23] L Zeng and MM Zhou. “Bromodomain: an acetyl-lysine binding domain”. In: {*FEBS*} *Letters* 513.1 (2002), pp. 124 –128.

[24] KL Yap and MM Zhou. “Structure and Mechanisms of Lysine Methylation Recognition by the Chromodomain in Gene Transcription”. In: *Biochemistry* 50.12 (2011), pp. 1966–1980.

[25] R Sanchez and MM Zhou. “The {PHD} finger: a versatile epigenome reader”. In: *Trends in Biochemical Sciences* 36.7 (2011), pp. 364 –372.

[26] The ENCODE Project Consortium. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489 (7414 2012), pp. 57–74.

[27] KS Zaret and JS Carroll. “Pioneer transcription factors: establishing competence for gene expression”. In: *Genes & Development* 25.21 (2011), pp. 2227–2241.

[28] AV Krivtsov and SA Armstrong. “MLL translocations, histone modifications and leukaemia stem-cell development”. In: *Nat Rev Cancer* 7 (11 2007), pp. 823–833.

[29] GJ Filion et al. “Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells”. In: *Cell* 143.2 (2010), pp. 212 –224.

[30] PV Kharchenko et al. “Comprehensive analysis of the chromatin landscape in Drosophila melanogaster”. In: *Nature* 471 (7339 2011), pp. 480–485.

[31] MI Arnone and EH Davidson. “The hardwiring of development: organization and function of genomic regulatory systems”. In: *Development* 124.10 (1997), pp. 1851–1864.

[32] S Istrail and EH Davidson. “Logic functions of the genomic cis-regulatory code”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.14 (2005), pp. 4954–4959.

[33] B Tolhuis et al. “Looping and Interaction between Hypersensitive Sites in the Active -globin Locus”. In: *Molecular Cell* 10.6 (2002), pp. 1453 –1465.

[34] W Deng et al. “Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor”. In: *Cell* 149.6 (2012), pp. 1233 –1244.

[35] JE Phillips-Cremins and VG Corces. “Chromatin Insulators: Linking Genome Organization to Cellular Function”. In: *Molecular Cell* 50.4 (2013), pp. 461 –474.

[36] VW Zhou, A Goren, and BE Bernstein. “Charting histone modifications and the functional organization of mammalian genomes”. In: *Nat Rev Genet* 12 (1 2011), pp. 7–18.

[37] B Li, M Carey, and JL Workman. “The Role of Chromatin during Transcription”. In: *Cell* 128.4 (2007), pp. 707 –719.

[38] BD Strahl and CD Allis. “The language of covalent histone modifications”. In: *Nature* 403 (6765 2000), pp. 41–45.

[39] SL Schreiber and BE Bernstein. “Signaling Network Model of Chromatin”. In: *Cell* 111.6 (2002), pp. 771 –778.

[40] Z Wang, M Gerstein, and M Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nat Rev Genet* 10 (1 2009), pp. 57–63.

[41] T Shiraki et al. “Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage”. In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15776–15781.

[42] R Kodzius et al. “CAGE: cap analysis of gene expression”. In: *Nat Meth* 3 (3 2006), pp. 211–222.

[43] A Barski et al. “High-Resolution Profiling of Histone Methylations in the Human Genome”. In: *Cell* 129.4 (2007), pp. 823 –837.

[44] DS Johnson et al. “Genome-Wide Mapping of in Vivo Protein-DNA Interactions”. In: *Science* 316.5830 (2007), pp. 1497–1502.

[45] G Robertson et al. “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing”. In: *Nat Meth* 4 (8 2007), pp. 651–657.

[46] B van Steensel and S Henikoff. “Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase”. In: *Nat Biotech* 18 (4 2000), pp. 424–428.

[47] N Plongthongkum, DH Diep, and K Zhang. “Advances in the profiling of DNA modifications: cytosine methylation and beyond”. In: *Nat Rev Genet* 15 (10 2014), pp. 647–661.

[48] E de Wit and W de Laat. “A decade of 3C technologies: insights into nuclear organization”. In: *Genes & Development* 26 (1 2012), pp. 11–24.

[49] MJ Fullwood et al. “An oestrogen-receptor-[agr]-bound human chromatin interactome”. In: *Nature* 462 (7269 2009), pp. 58–64.

[50] K Sachs et al. “Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data”. In: *Science* 308.5721 (2005), pp. 523–529.

[51] A Margolin et al. “ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context”. In: *BMC Bioinformatics* 7.Suppl 1 (2006), S7.

[52] D Juan, F Pazos, and A Valencia. “High-confidence prediction of global interactomes based on genome-wide coevolutionary networks”. In: *Proceedings of the National Academy of Sciences* 105.3 (2008), pp. 934–939.

[53] JG van Bemmel et al. “A Network Model of the Molecular Organization of Chromatin in Drosophila”. In: *Molecular Cell* 49.4 (2013), pp. 759 –771.

[54] J Lasserre, HR Chung, and M Vingron. “Finding associations among histone modifications using sparse partial correlation networks”. In: *PLoS Comput Biol* 9.9 (2013), e1003168.

[55] J Zhou and OG Troyanskaya. “Global Quantitative Modeling of Chromatin Factor Interactions”. In: *PLoS Comput Biol* 10.3 (2014), e1003525.

[56] MB Eisen et al. “Cluster analysis and display of genome-wide expression patterns”. In: *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868.

[57] A Butte and I Kohane. “Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements”. In: *Pac Symp Biocomput* 5 (2000), pp. 418–429.

[58] SL Lauritzen. *Graphical Models.* Oxford: Clarendon Press, 1996, pp. 137–156.

[59] AP Dempster. “Covariance Selection”. In: *Biometrics* 28.1 (1972), pp. 157–175.

[60] N Meinshausen and P Bühlmann. “High-dimensional graphs and variable selection with the Lasso”. In: *The Annals of Statistics* 34.3 (2006), pp. 1436–1462.

[61] J Friedman, T Hastie, and R Tibshirani. “Sparse inverse covariance estimation with the graphical lasso”. In: *Biostatistics* 9.3 (2008), pp. 432–441.

[62] F Markowetz and R Spang. “Inferring cellular networks - a review”. In: *BMC Bioinformatics* 8.Suppl 6 (2007), S5.

[63] M Weigt et al. “Identification of direct residue contacts in proteinprotein interaction by message passing”. In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.

[64] I Tur, A Roverato, and R Castelo. “Mapping eQTL networks with mixed graphical Markov models”. In: *Genetics* 198.4 (2014), pp. 1377–1393.

[65] Zou H and Hastie T. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2 2005), 301320.

[66] T Murali et al. “DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila”. In: *Nucleic Acids Research* 39.suppl 1 (2011), pp. D736–D743.

[67] KD Pruitt et al. “efSeq: an update on mammalian reference sequences”. In: *Nucleic Acids Research* 42.D1 (2014), pp. D756–D763.

[68] B Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10.3 (2009), R25.

[69] V Kumar et al. “Uniform, optimal signal processing of mapped deep-sequencing data”. In: *Nat Biotech* 31 (7 2013), pp. 615–622.

[70] X Dong et al. “Modeling gene expression using chromatin features in various cellular contexts”. In: *Genome Biology* 13.9 (2012), R53.

[71] R Karlić et al. “Histone modification levels are predictive for gene expression”. In: *Proceedings of the National Academy of Sciences* 107.7 (2010), pp. 2926–2931.

[72] J Persson and K Ekwall. “Chd1 remodelers maintain open chromatin and regulate the epigenetics of differentiation”. In: *Experimental Cell Research* 316.8 (2010), pp. 1316 –1323.

[73] G Hu et al. “H2A.Z Facilitates Access of Active and Repressive Complexes to Chromatin in Embryonic Stem Cell Self-Renewal and Differentiation”. In: *Cell Stem Cell* 12.2 (2013), pp. 180 –192.

[74] B Wittschieben et al. “A Novel Histone Acetyltransferase Is an Integral Subunit of Elongating {RNA} Polymerase {II} Holoenzyme”. In: *Molecular Cell* 4.1 (1999), pp. 123 –128.

[75] JH Kim, WS Lane, and D Reinberg. “Human Elongator facilitates RNA polymerase II transcription through chromatin”. In: *Proceedings of the National Academy of Sciences* 99.3 (2002), pp. 1241–1246.

[76] B Czermin et al. “Drosophila Enhancer of Zeste/ESC Complexes Have a Histone {H3} Methyltransferase Activity that Marks Chromosomal Polycomb Sites”. In: *Cell* 111.2 (2002), pp. 185 –196.

[77] J Müller et al. “Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex”. In: *Cell* 111.2 (2002), pp. 197 –208.

[78] A Kuzmichev et al. “Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein”. In: *Genes & Development* 16.22 (2002), pp. 2893–2905.

[79] R Cao et al. “Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing”. In: *Science* 298.5595 (2002), pp. 1039–1043.

[80] KH Hansen et al. “A model for transmission of the H3K27me3 epigenetic mark”. In: *Nat Cell Biol* 10 (11 2008), pp. 1291–1300.

[81] R Margueron et al. “Role of the polycomb protein EED in the propagation of repressive histone marks”. In: *Nature* 461 (7265 2009), pp. 762–767.

[82] AR Bataille et al. “A Universal {RNA} Polymerase {II} {CTD} Cycle Is Orchestrated by Complex Interplays between Kinase, Phosphatase, and Isomerase Enzymes along Genes”. In: *Molecular Cell* 45.2 (2012), pp. 158 –170.

[83] KO Kizer et al. “A Novel Domain in Set2 Mediates RNA Polymerase II Interaction and Couples Histone H3 K36 Methylation with Transcript Elongation”. In: *Molecular and Cellular Biology* 25.8 (2005), pp. 3305–3316.

[84] Y Li et al. “The Target of the NSD Family of Histone Lysine Methyltransferases Depends on the Nature of the Substrate”. In: *Journal of Biological Chemistry* 284.49 (2009), pp. 34283–34295.

[85] K Nimura et al. “A histone H3 lysine 36 trimethyltransferase links Nkx2-5 to Wolf-Hirschhorn syndrome”. In: *Nature* 460 (7252 2009), pp. 287 –291.

[86] F Forneris et al. “Histone demethylation catalysed by {LSD1} is a flavin-dependent oxidative process”. In: *FEBS Letters* 579.10 (2005), pp. 2203 –2207.

[87] Y Shi et al. “Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog {LSD1}”. In: *Cell* 119.7 (2004), pp. 941 –953.

[88] O Binda et al. “Trimethylation of histone H3 lysine 4 impairs methylation of histone H3 lysine 9”. In: *Epigenetics* 5.8 (2010), pp. 767–775.

[89] C He et al. “The Methyltransferase NSD3 Has Chromatin-binding Motifs, PHD5-C5HCH, That Are Distinct from Other NSD (Nuclear Receptor SET Domain) Family Members in Their Histone H3 Recognition”. In: *Journal of Biological Chemistry* 288.7 (2013), pp. 4692–4703.

[90] SK Kim et al. “Human Histone H3K79 Methyltransferase DOT1L Methyltransferase Binds Actively Transcribing RNA Polymerase II to Regulate Gene Expression”. In: *Journal of Biological Chemistry* 287.47 (2012), pp. 39698–39709.

[91] F Frederiks et al. “Nonprocessive methylation by Dot1 leads to functional redundancy of histone H3K79 methylation states”. In: *Nat Struct Mol Biol* 15 (6 2008), pp. 550–557.

[92] M Lachner et al. “Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins”. In: *Nature* 410 (6824 2001), pp. 116–120.

[93] Y Li et al. “The Histone Modifications Governing TFF1 Transcription Mediated by Estrogen Receptor”. In: *Journal of Biological Chemistry* 286.16 (2011), pp. 13925–13936.

[94] Q Jin et al. “Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation”. In: *The EMBO Journal* 30.2 (2010), pp. 249–262.

[95] H Yuan and R Marmorstein. “Histone acetyltransferases: Rising ancient counterparts to protein kinases”. In: *Biopolymers* 99.2 (2013), pp. 98–111.

[96] RJ Sims et al. “Human but Not Yeast CHD1 Binds Directly and Selectively to Histone H3 Methylated at Lysine 4 via Its Tandem Chromodomains”. In: *Journal of Biological Chemistry* 280.51 (2005), pp. 41789–41792.

[97] JF Flanagan et al. “Double chromodomains cooperate to recognize the methylated histone H3 tail”. In: *Nature* 438 (7071 2005), pp. 1181–1185.

[98] L Bugga et al. “The Drosophila melanogaster CHD1 Chromatin Remodeling Factor Modulates Global Chromosome Structure and Counteracts HP1a and H3K9me2”. In: *PLoS ONE* 8.3 (2013), e59496.

[99] W Feng et al. “PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation”. In: *Nat Struct Mol Biol* 17 (4 2010), pp. 445–450.

[100] JR Horton et al. “Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases”. In: *Nat Struct Mol Biol* 17 (1 2010), pp. 38–43.

[101] K Fortschegger et al. “PHF8 Targets Histone Methylation and RNA Polymerase II To Activate Transcription”. In: *Molecular and Cellular Biology* 30.13 (2010), pp. 3286–3298.

[102] M Vermeulen et al. “Quantitative Interaction Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers”. In: *Cell* 142.6 (2010), pp. 967 –980.

[103] HH Qi et al. “Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development”. In: *Nature* 466 (7305 2010), pp. 503–507.

[104] WJ Shia, B Li, and JL Workman. “SAS-mediated acetylation of histone H4 Lys 16 is required for H2A.Z incorporation at subtelomeric regions in Saccharomyces cerevisiae”. In: *Genes & Development* 20 (18 2006), pp. 2507–2512.

[105] JJ Pesavento et al. “Certain and Progressive Methylation of Histone H4 at Lysine 20 during the Cell Cycle”. In: *Molecular and Cellular Biology* 28.1 (2008), pp. 468–486.

[106] T Klymenko et al. “A Polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities”. In: *Genes & Development* 20.9 (2006), pp. 1110–1122.

[107] J Ernst and M Kellis. “ChromHMM: automating chromatin-state discovery and characterization”. In: *Nat Meth* 9 (3 2012), pp. 215–216.

[108] J Ernst and M Kellis. “Discovery and characterization of chromatin states for systematic annotation of the human genome”. In: *Nat Biotech* 28 (8 2010), pp. 817–825.

Appendices

APPENDIX *A*

Supplementary Figures

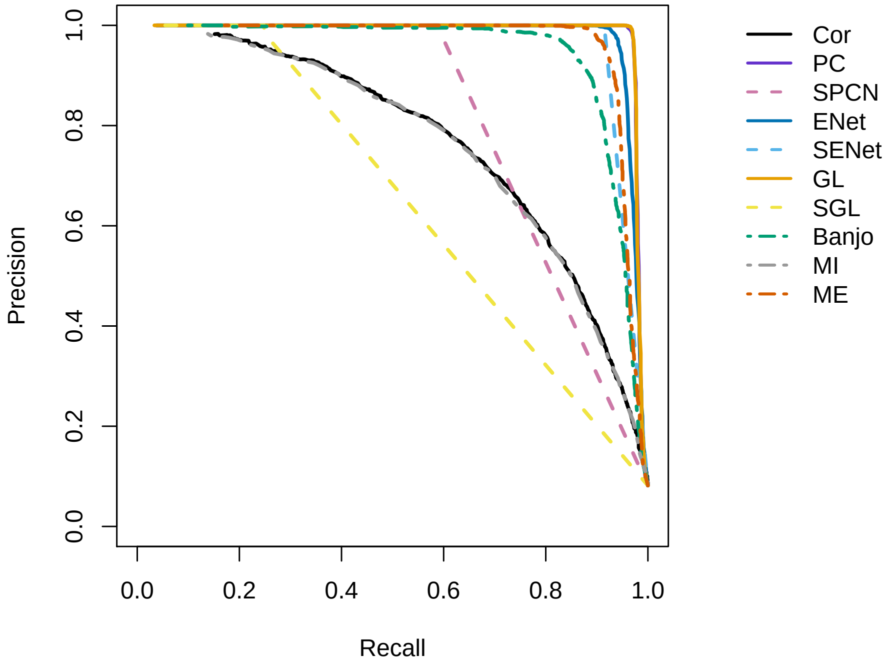

Figure A.1.: *The average precision-recall curves over 25 simulated networks with 50 nodes, 100 edges and 10.000 observations. Solid and dashed lines indicate methods that are run on continuous data and their corresponding sparse version, respectively. Dotted-dashed lines indicate methods run on discrete data.*

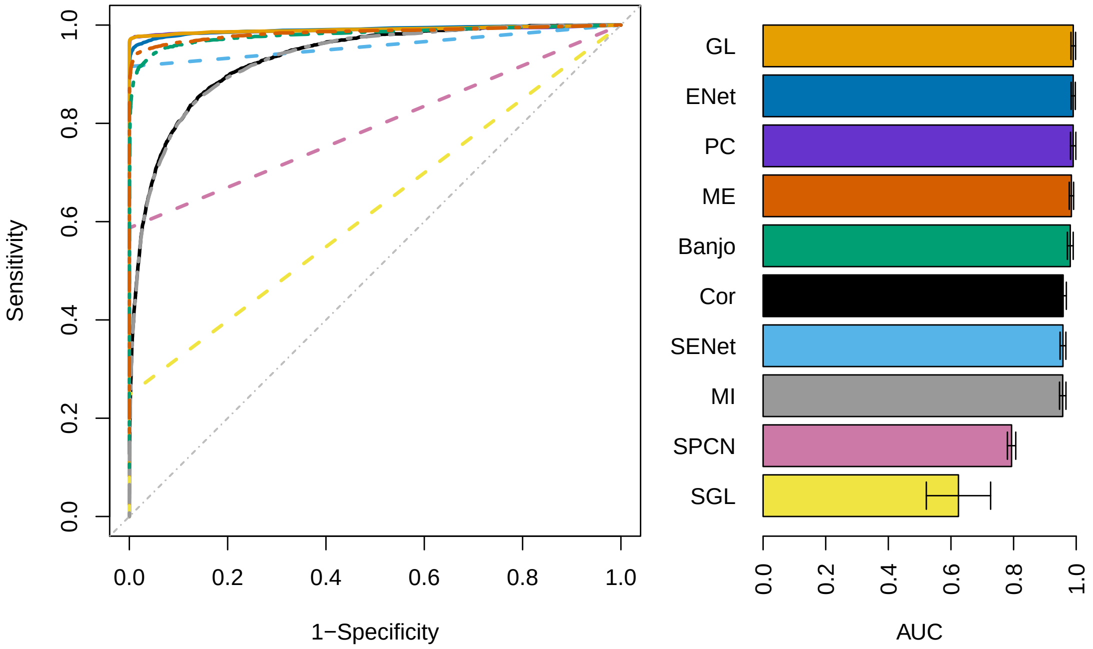

Figure A.2.: *The average ROC curves (left plot) over 25 random networks with 50 nodes, 100 edges and 10.000 observations. The plot on the right indicates the corresponding average AUC and the standard deviation over the 25 random networks. Solid and dashed lines indicate methods that are run on continuous data and their corresponding sparse version, respectively. Dotted-dashed lines indicate methods run on discrete data.*

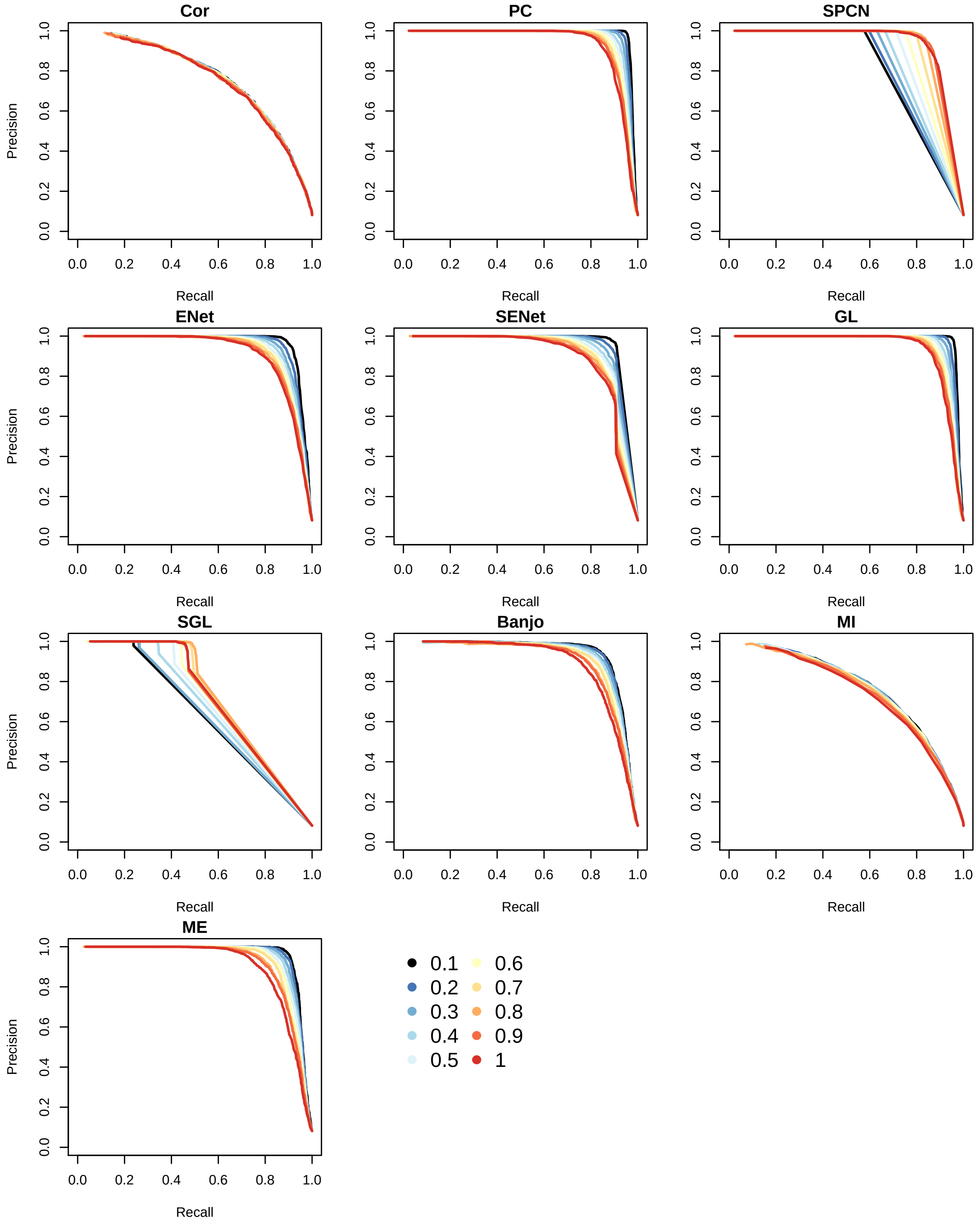

Figure A.3.: *Average precision recall curves for different percentages of univariate Gaussian noise added to the data. Curves are averaged over 25 simulated networks. The simulated networks have 50 nodes and 100 edges.*

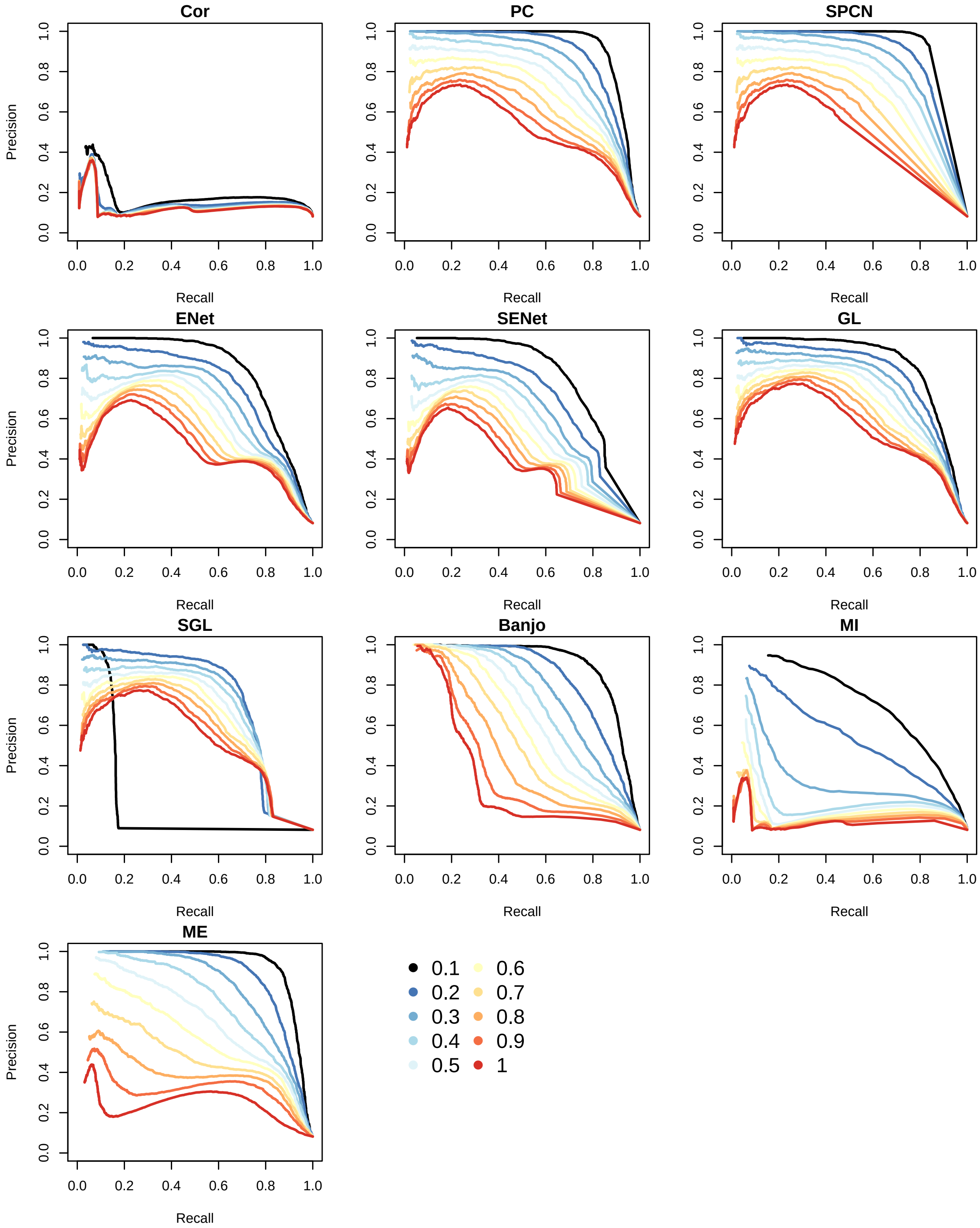

Figure A.4.: *Average precision recall curves for different percentages of multivariate Gaussian noise added to the data. Curves are averaged over* 25 *simulated networks. The simulated networks have* 50 *nodes and* 100 *edges.*

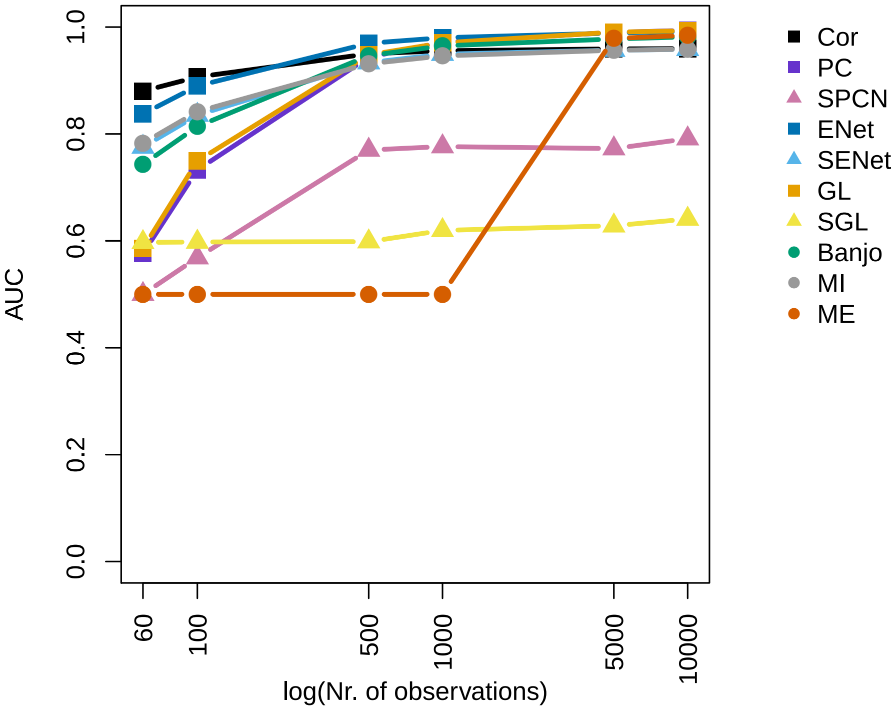

Figure A.5.: *The average AUC with increasing number of corresponding observations. The performance of the different network reconstruction methods is averaged over* 25 *simulated networks each having* 50 *nodes and* 100 *edges.*

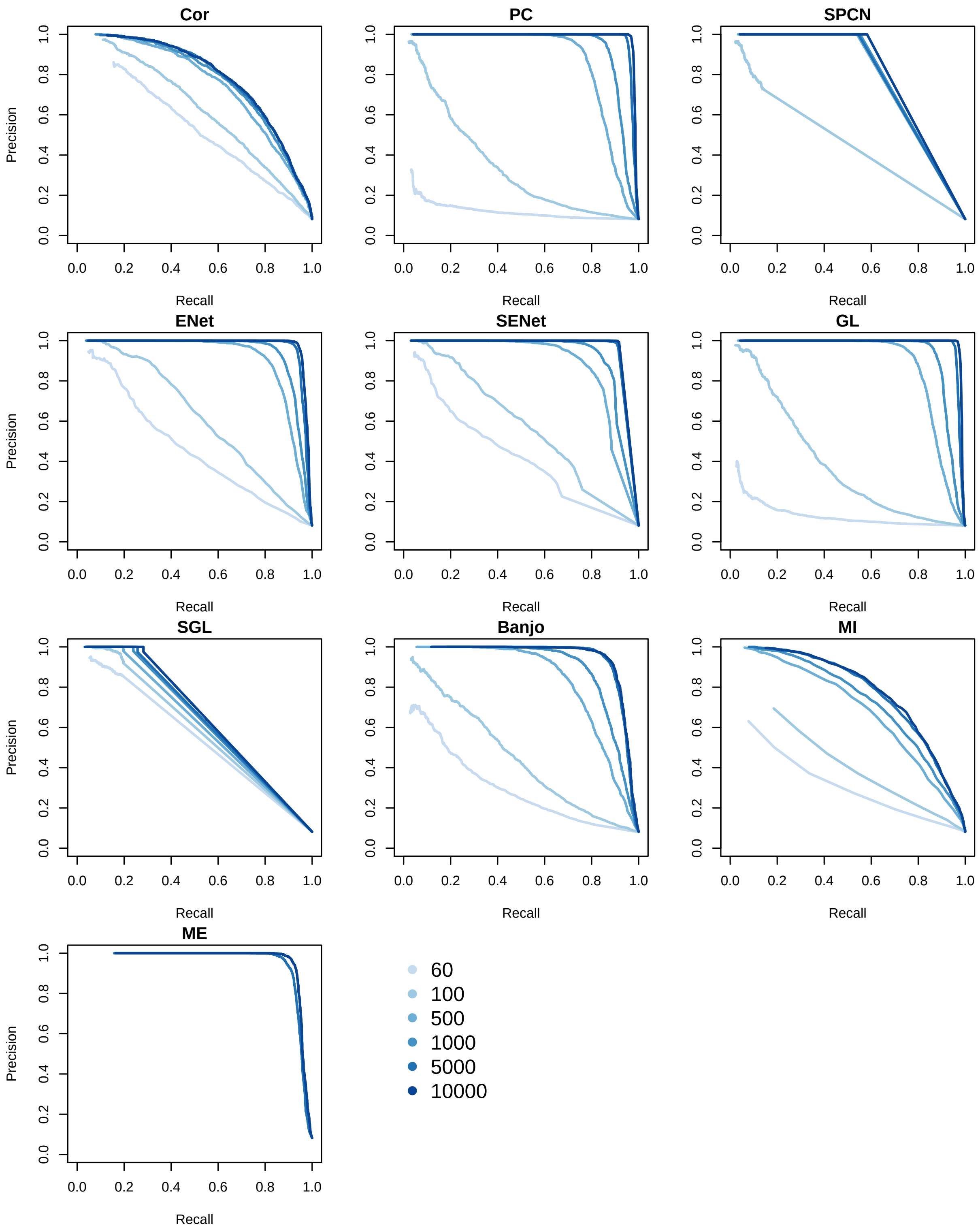

Figure A.6.: *Average precision recall curves for different numbers of observations averaged over* 25 *simulated networks. The simulated networks have* 50 *nodes and* 100 *edges.*

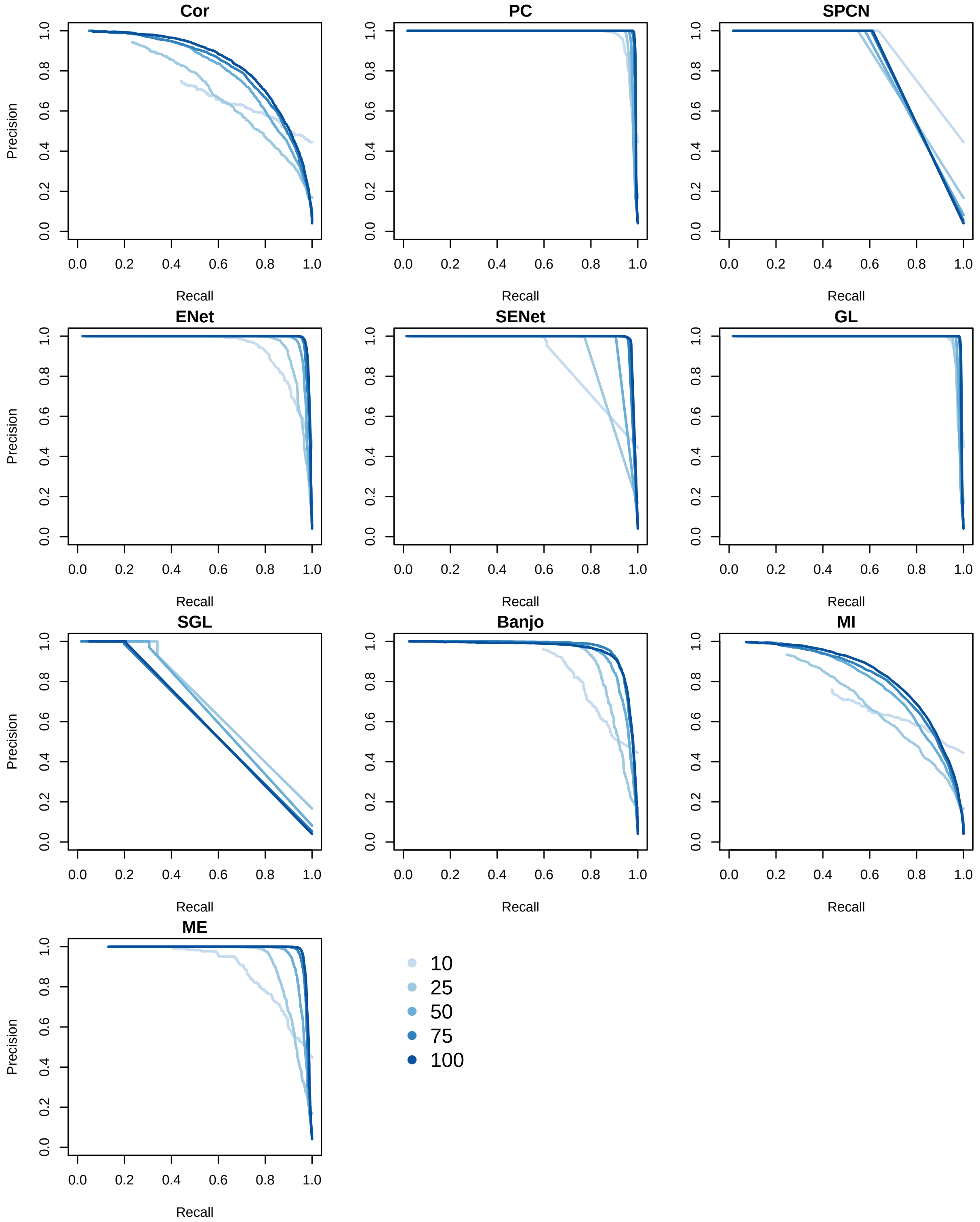

Figure A.7.: *Average precision recall curves for different numbers of variables in the simulated network averaged over* 25 *networks. The methods were trained on* 10.000 *observations*

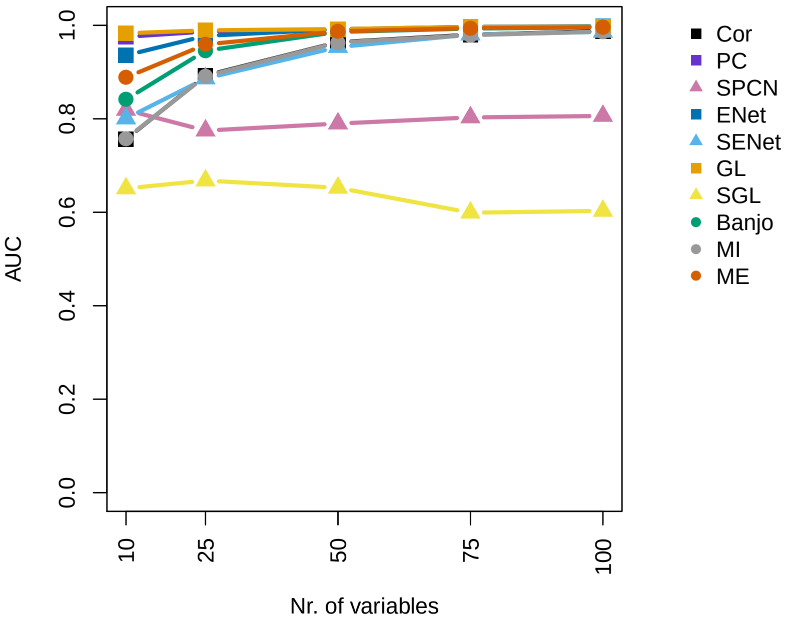

Figure A.8.: *The average AUC over* 25 *simulated networks with increasing number of variables in the simulated networks. The networks were trained on* 10.000 *observations*

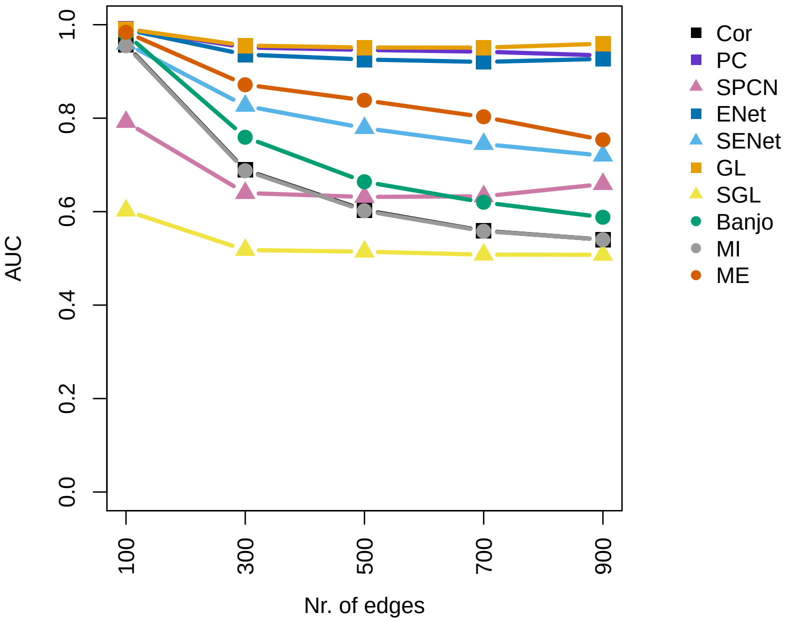

Figure A.9.: *The average AUC with increasing graph density/number of edges. The performance of the different network reconstruction methods is averaged over* 25 *simulated networks having* 50 *nodes and based on* 10.000 *observations.*

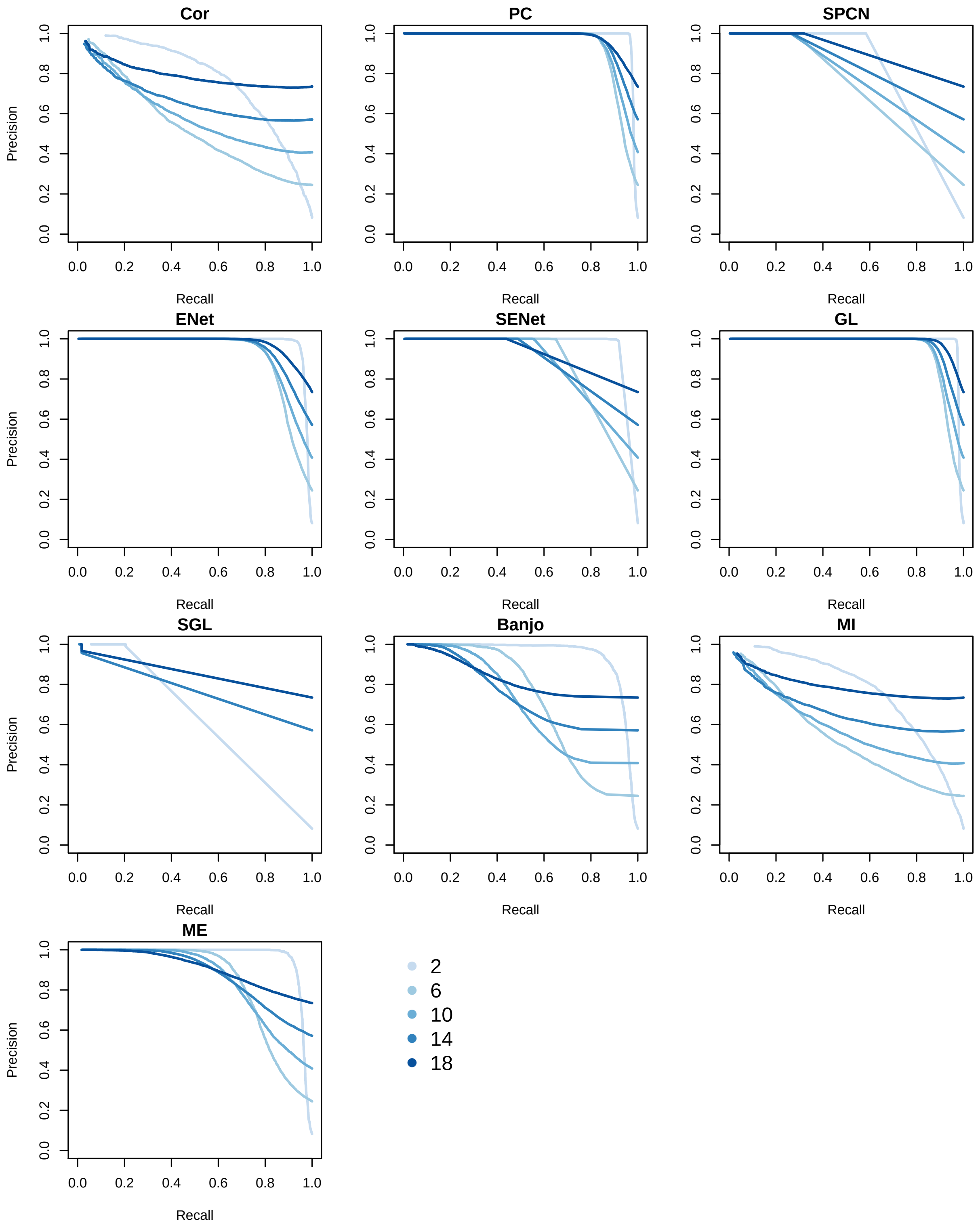

Figure A.10.: *Precision recall plots for each methods on simulated networks with between* 100 *(node degree* 2*) and* 900 *(node degree* 18*) edges. The performance of the different network reconstruction methods is averaged over* 25 *simulated networks having* 50 *nodes and based on* 10.000 *observations.*

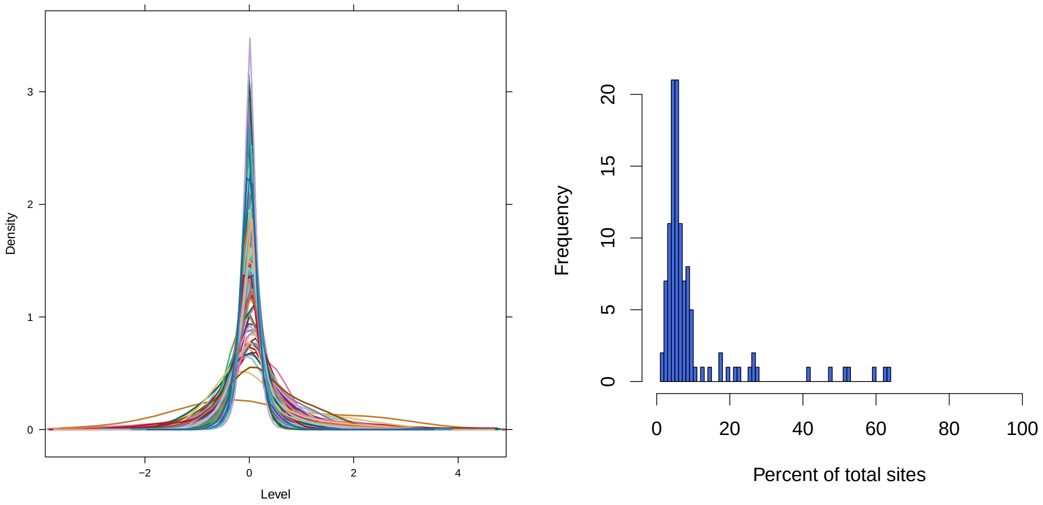

Figure A.11.: *Density curves and peak counts of the HMs and CMs in Drosophila.*

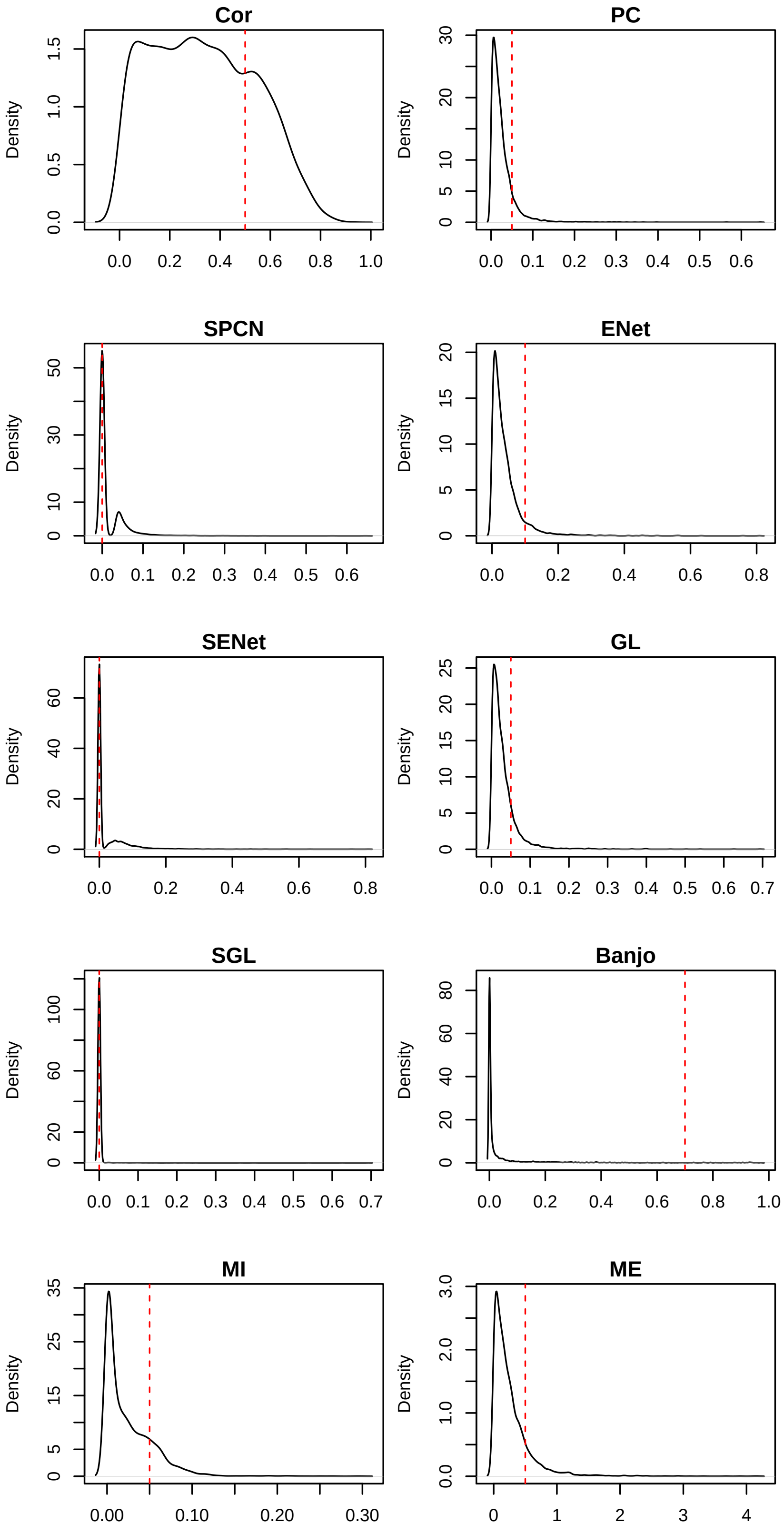

Figure A.12.: *Distribution of the absolute edge scores for each network reconstruction method applied to Drosophila DamID data. The red dashed line indicates the arbitrarily chosen cutoff to call edges in the network.*

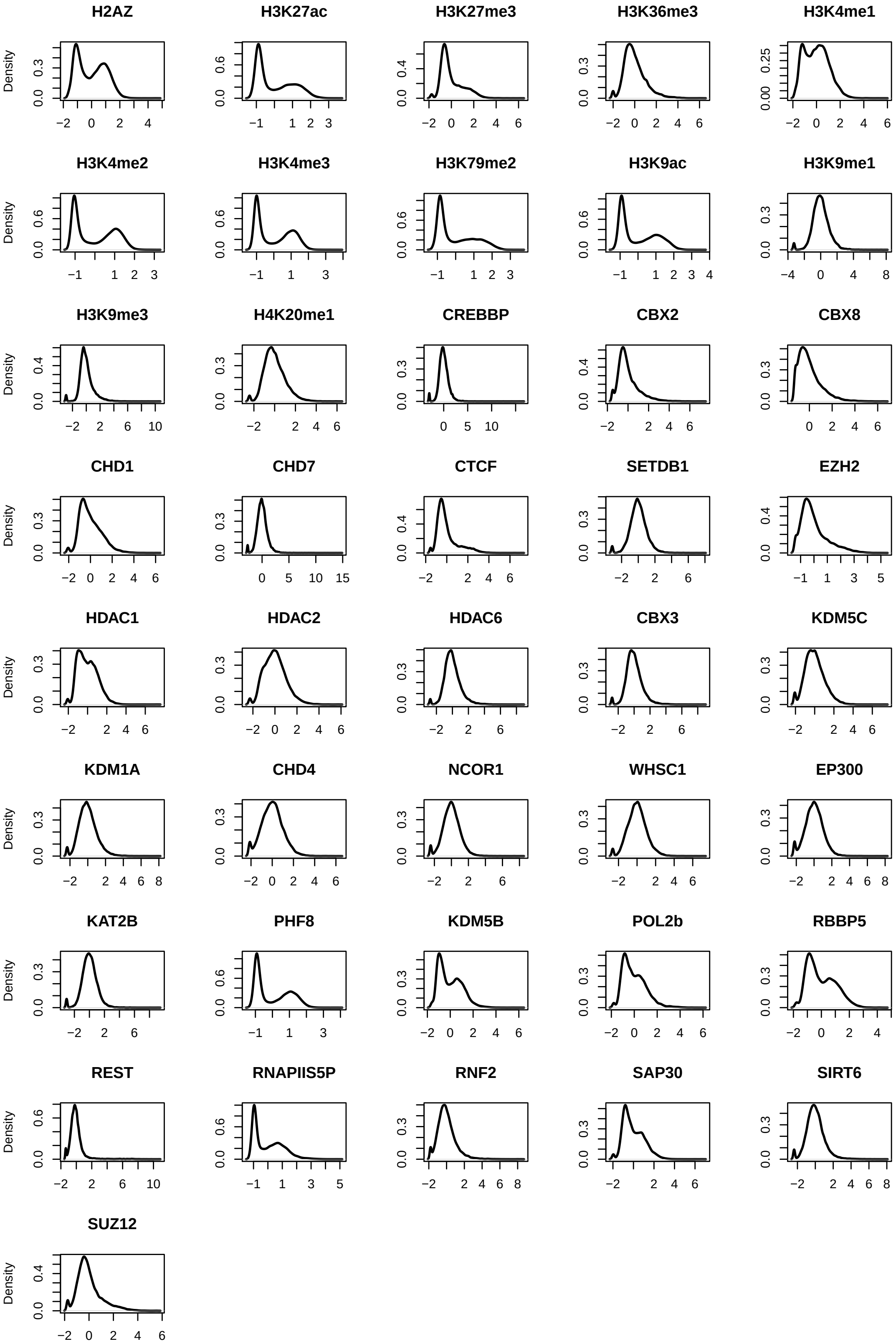

Figure A.13.: *Density curves of the normalized HM and CM signals in the human K562 cell line.*

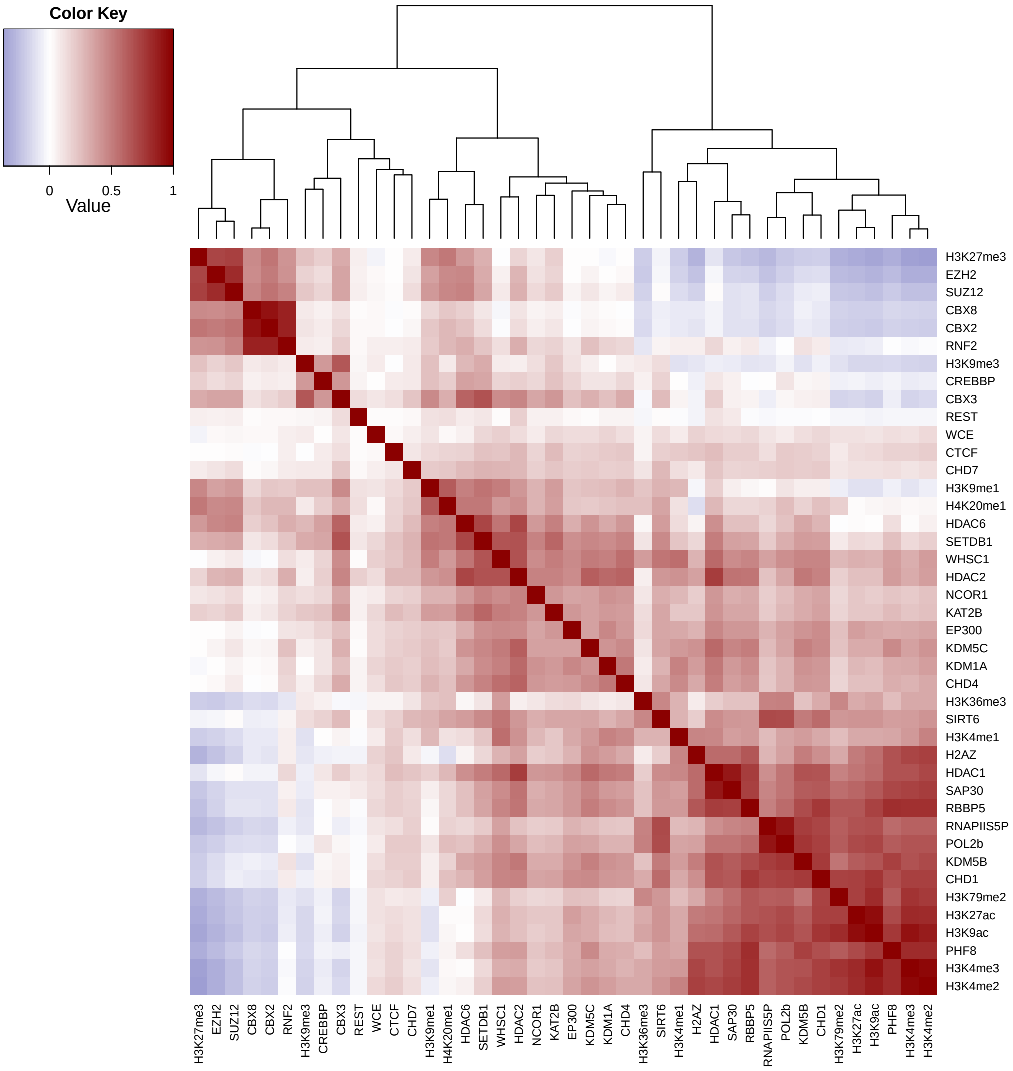

Figure A.14.: *Heatmap of the Pearson correlation coefficients as indicated by the color key between HM and CM signals in the human K562 cell line.*

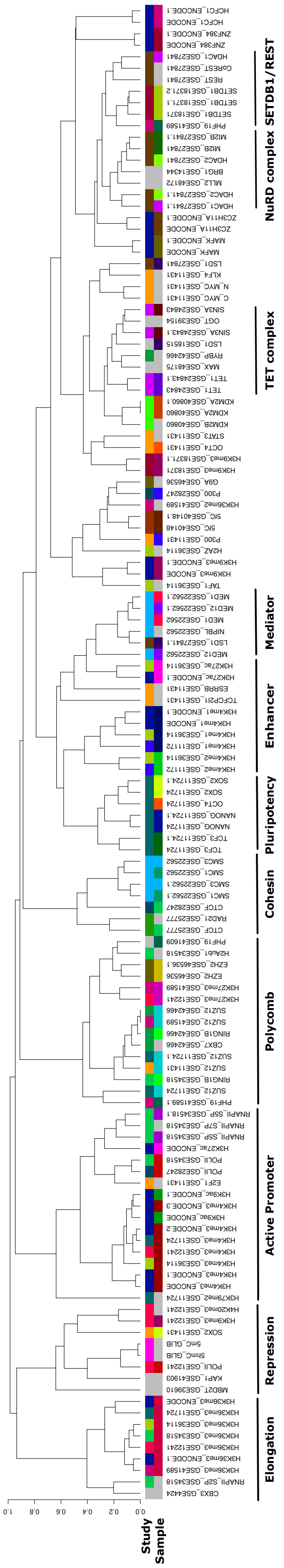

Figure A.15.: *Hierarchical clustering of all available experiments including all replicates on HMs, chromatin-binding proteins and cytosine modifications. The distance measure is based on genome-wide correlation. The upper color label indicates groups of experiments from the same study and the lower color label indicates groups of replicates. Studies or replicates where only one experiment is available are indicated in grey.*

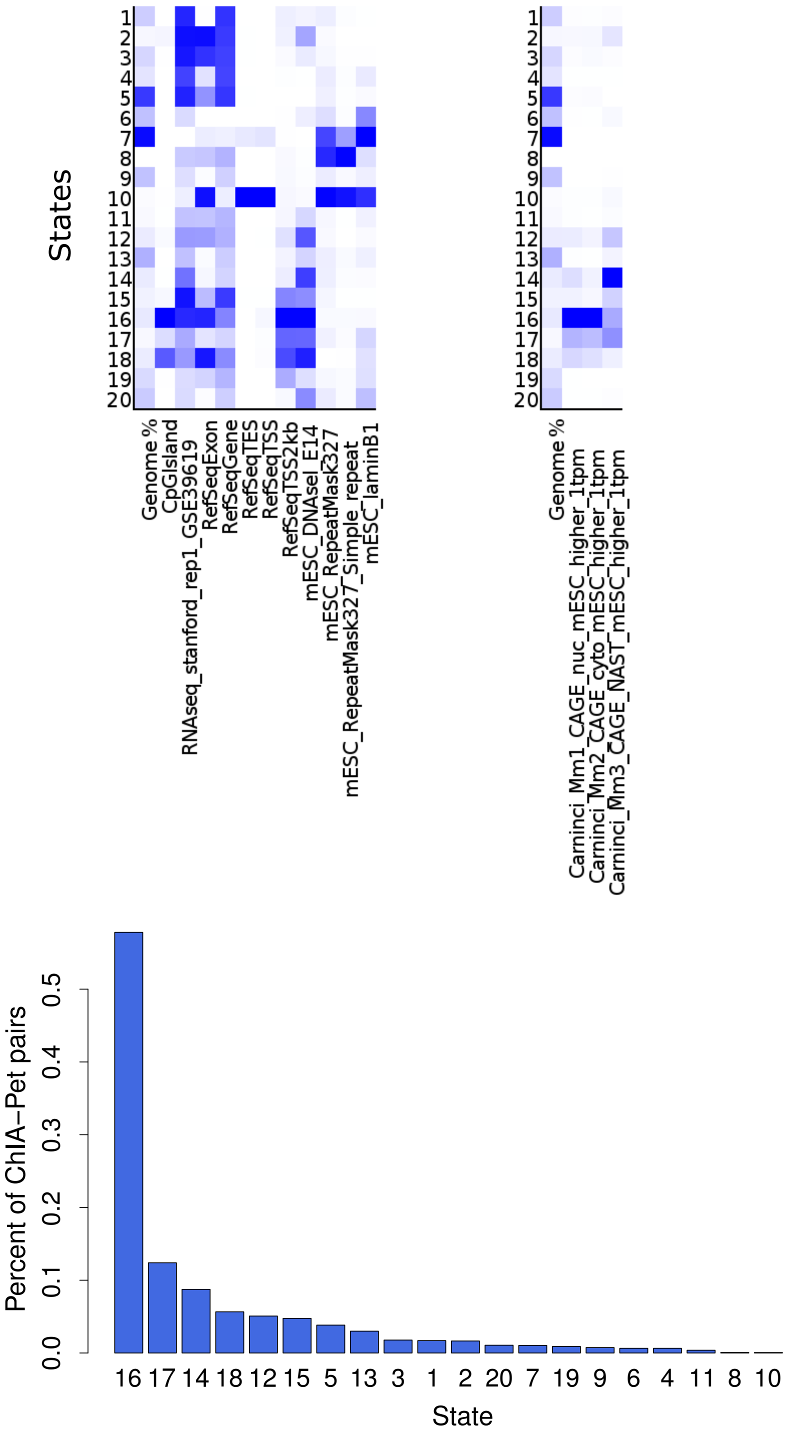

Figure A.16.: *The heatmaps summarize the enrichment (indicated by color intensity from white (no enrichment) to blue (high enrichment)) for known annotations including CpG islands, RefSeq transcript annotations and repeat regions across the chromatin states. Further, enrichment for other cell-type specific data (i.e. RNA-Seq, Dnase hypersensitivity, LaminB1) is shown. The figures were provided by our collaborators. Last plot shows the percent of all regions participating in a ChIA-Pet pair coinciding with a particular chromatin state.*

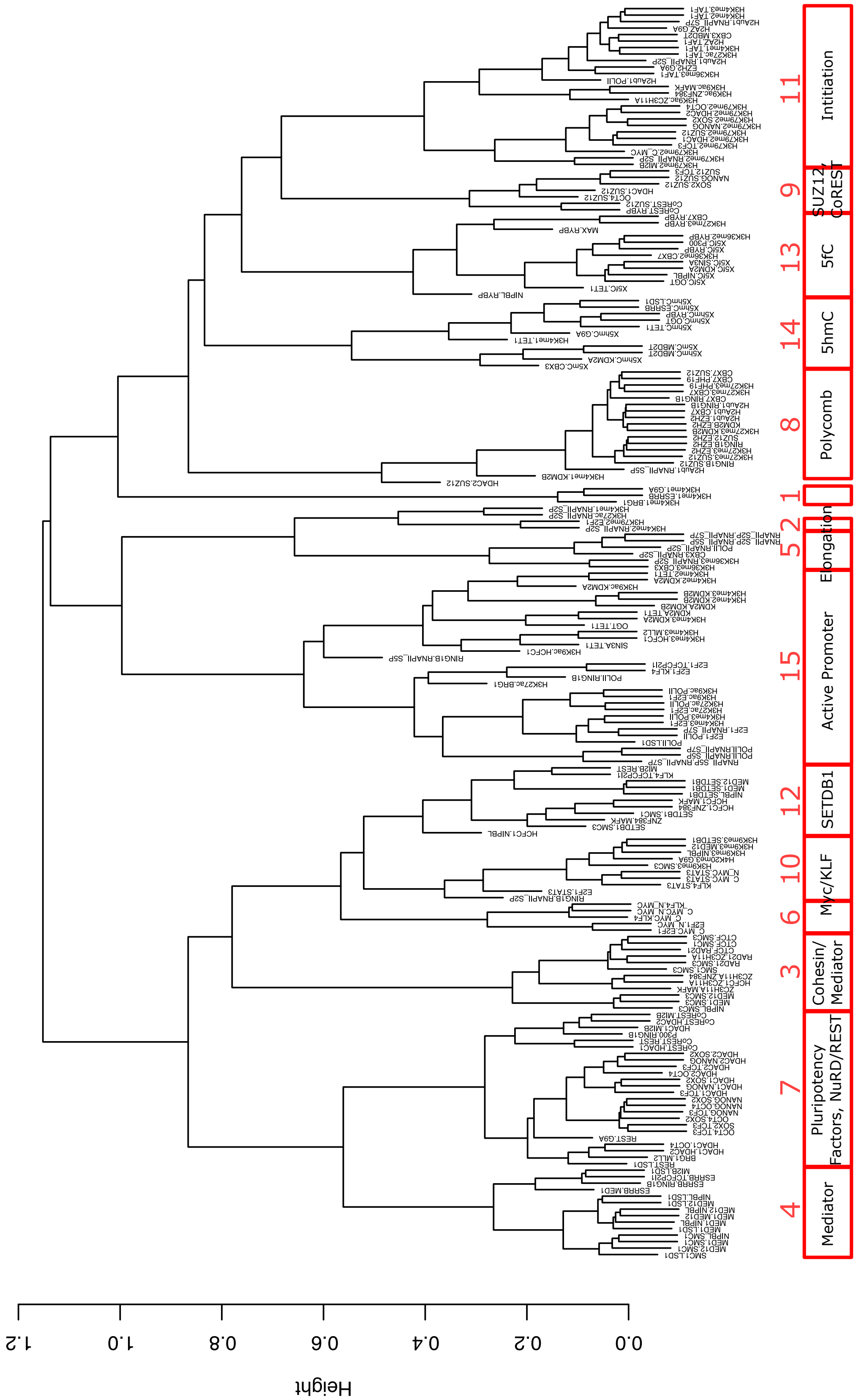

Figure A.17.: *Hierarchical clustering of the chromatin state-specific interaction based on their enrichment profiles across chromatin-states. Red boxes indicate detected sub-networks with manual functional annotation based on the participating chromatin modifications and chromatin-related proteins. The figure was provided by our collaborators.*

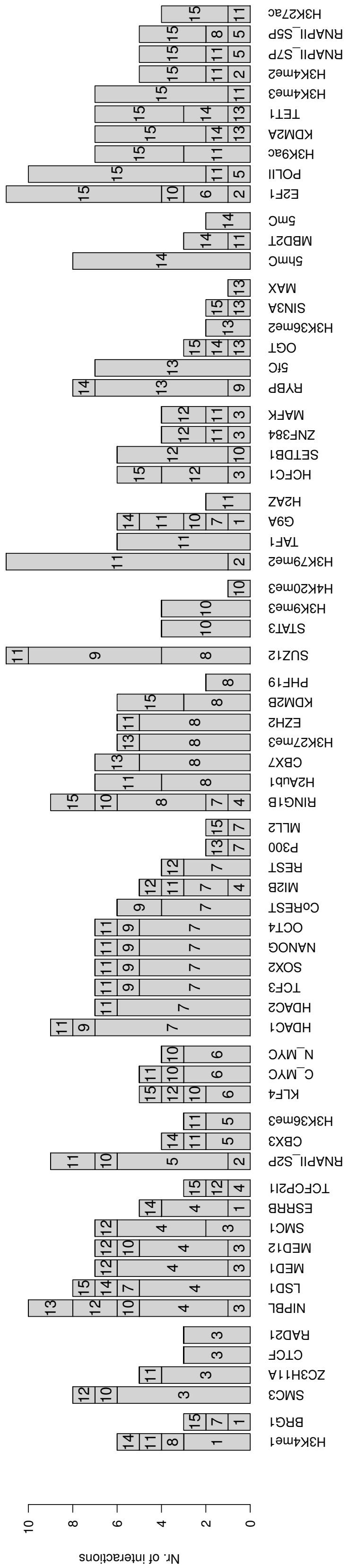

Figure A.18.: *Number of interaction that a protein takes part in split up across the different subnetworks.*

APPENDIX B

Supplementary Tables

Table B.1.: *Accession numbers of the ChIP-Seq experiments of HMs and CMs used to study the genome-wide interactions in humans.*

HM/CM	K562 - GEO accession ID	H1 - GEO accession ID
WCE	GSM733780, GSM831024, GSM831023, GSM831022	GSM831043, GSM831044
RBBP5	GSM831014, GSM831013, GSM831012	GSM831039, GSM831038
SAP30	GSM831019, GSM831018	GSM831040
PHF8	GSM831009, GSM831008	GSM831037
HDAC6	GSM830998, GSM830997	GSM831030
HDAC2	GSM830996, GSM830995	GSM831029
CHD1	GSM830989, GSM830988	GSM831026, GSM831025
CTCF	GSM733719	GSM646335, GSM646334
SUZ12	GSM831021	GSM831042
SIRT6	GSM831020	GSM831041
P300	GSM831006	GSM831036
JARID1C	GSM831000	GSM831036
EZH2	GSM830992	GSM831028
CHD7	GSM830990	GSM831027
HDAC1	GSM830994, GSM830993	
PLU1	GSM831011, GSM831010	
POL2b	GSM733643	
RNF2	GSM831017	
RNAPIIS5P	GSM831016	
REST	GSM831015	
PCAF	GSM831007	
NSD2	GSM831005	
NCOR	GSM831004	
MI2	GSM831003	
LSD1	GSM831002	
CBP	GSM831001	
HP1G	GSM830999	
ESET	GSM830991	
CBX8	GSM830987	
CBX2	GSM830986	
H3K4me3	GSM733680	GSM409308, GSM469971, GSM433170, GSM646346, GSM646345, GSM432392, GSM410808, GSM605315, GSM593365
H3K4me1	GSM733692	GSM409307, GSM466739, GSM433177, GSM646342, GSM646341, GSM605312, GSM434762
H3K27me3	GSM733658	GSM434776, GSM466734, GSM433167, GSM646338, GSM646337, GSM605308, GSM428295
H3K9ac	GSM733778	GSM434785, GSM433171, GSM646348, GSM646347, GSM410807, GSM605323
H3K9me3	GSM733776	GSM433174, GSM605328, GSM605327, GSM605325, GSM450266, GSM428291
H3K4me2	GSM733651	GSM646344, GSM646343, GSM602261, GSM602260
H4K20me1	GSM733675	GSM646350, GSM646349, GSM605329
H3K27ac	GSM733656	GSM663427, GSM466732, GSM646336
H3K79me2	GSM733653	GSM605322, GSM605321
H2A.Z	GSM733786	GSM1003579
H3K9me1	GSM733777	
H3K36me3	GSM733714	

Table B.2.: *Accession numbers of the experiments used to study the genome-wide interactions in mouse embryonic stem cells.*

Data	GEO accession ID
DNA control	GSE11724, GSE12241, GSE14344, GSE18371, GSE18515, GSE22562, GSE24030, GSE24843, GSE27841, GSE28247, GSE34518, GSE28682, GSE32218-ENCODE, GSE36030-ENCODE, GSE36114, GSE39154, GSE39610, GSE40148, GSE40860, GSE41589, GSE41589, GSE41609, GSE41903, GSE42466, GSE44242, GSE46536, GSE48172, GSE48175, GSE40810
H3K36me3	GSE11724, GSE12241, GSE31039-ENCODE, GSE34518, GSE36114, GSE41589
H3K4me3	GSE11724, GSE12241, GSE31039-ENCODE, GSE32218-ENCODE, GSE36114
H3K9me3	GSE12241, GSE18371, GSE32218-ENCODE
H3K4me1	GSE11172, GSE31039-ENCODE, GSE36114
H3K27me3	GSE12241, GSE36114, GSE41589
H3K27ac	GSE31039-ENCODE, GSE36114
H3K4me2	GSE11172, GSE36114
H3K9ac	GSE31039-ENCODE
H3K79me2	GSE11724
H4K20me3	GSE12241
H3K36me2	GSE41589
H2Aub1	GSE34518
H2AZ	GSE36114
5hmC	GSE28682, GSE40810
5mC	GSE28682
5fC	GSE40148
SUZ12	GSE11431, GSE11724, GSE41589, GSE42466
EP300	GSE11431, GSE28247
RING1B	GSE34518, GSE42466
LSD1	GSE18515, GSE27841
PHF19	GSE41589, GSE41609
HCFC1	GSE36030-ENCODE
BRG1	GSE14344
SETDB1	GSE18371
MED12	GSE22562
MED1	GSE22562
NIPBL	GSE22562
SMC1	GSE22562
SMC3	GSE22562
RAD21	GSE24030
SIN3A	GSE24843
TET1	GSE24843
HDAC1	GSE27841
HDAC2	GSE27841
MI2B	GSE27841
REST	GSE27841
CoREST	GSE27841
TAF1	GSE36114
OGT	GSE39154
MBD2T	GSE39610
KDM2A	GSE40860
KDM2B	GSE40860
KAP1	GSE41903
CBX7	GSE42466
RYBP	GSE42466
CBX3	GSE44242
EZH2	GSE46536
G9A	GSE46536
MLL2	GSE48172
TCF3	GSE11724
RPOLII	GSE12241, GSE28247
RNAPII_8WG16	GSE34518
RNAPII_S2P	GSE34518
RNAPII_S5P	GSE34518
RNAPII_S7P	GSE34518
CTCF	GSE11431, GSE24030, GSE28247
NANOG	GSE11431, GSE11724
OCT4	GSE11431, GSE11724
SOX2	GSE11431, GSE11724
MAFK	GSE36030-ENCODE
ZC3H11A	GSE36030-ENCODE
ZNF384	GSE36030-ENCODE
E2F1	GSE11431
ESRRB	GSE11431
KLF4	GSE11431
N_MYC	GSE11431
C_MYC	GSE11431
STAT3	GSE11431
TCFCP2L1	GSE11431
MAX	GSE48175

Table B.3.: *Transition probabilities between chromatin states. Each entry gives the percent of segments annotated with the chromatin state indicated in the row where a transition to the chromatin state in the column follows or preceded. Only entries larger than 0.01 are shown.*

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	0.64	0.06	0.01	0.08	0.02										0.14	0.02	0.01			
2	0.19	0.23	0.34	0.01	0.04							0.01	0.01	0.05	0.08					
3	0.01	0.10	0.41	0.04	0.36							0.01	0.02							
4	0.05		0.04	0.82	0.07															
5	0.01		0.20	0.05	0.71															
6						0.88	0.04			0.02			0.01	0.01						
7							0.98			0.01										
8						0.01		0.17	0.52		0.04	0.04	0.07						0.01	0.12
9								0.04	0.65		0.05	0.04	0.10						0.02	0.09
10						0.17	0.57			0.21							0.01			0.02
11			0.02	0.01				0.03	0.41		0.15	0.16	0.06						0.02	0.11
12			0.02					0.01	0.14		0.08	0.18	0.34	0.02		0.02	0.04	0.02	0.01	0.10
13			0.02						0.15		0.01	0.12	0.52	0.08			0.02			0.04
14	0.03	0.06	0.02			0.05			0.01			0.03	0.43	0.23	0.03	0.03	0.05			0.02
15	0.37	0.07			0.01									0.02	0.15	0.34	0.02			
16	0.04											0.02	0.02	0.02	0.28	0.13	0.40	0.06		
17	0.02			0.02	0.01	0.01	0.01		0.03			0.04	0.06	0.03	0.01	0.42	0.21	0.04	0.01	0.04
18												0.02	0.01			0.07	0.04	0.20	0.61	
19									0.05				0.01					0.31	0.58	0.02
20			0.02	0.02	0.02	0.01	0.03	0.03	0.29		0.04	0.08	0.08				0.03		0.02	0.30

List of Abbreviations

5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylation
5mC	5-methylcytosine
A	Argenine
ac	Acetylation
ATP	Adenosine triphosphate
AUC	area under the receiver operating characteristic
bp	base pair
CAGE	Cap Analysis Gene Expression
DNMT	DNA methyltransferase
ENet	Elastic Net
hESC	human embryonic stem cell
HM	Histone modification
HMM	Hidden Markov Model
HMM	hidden markov model
K	Lysine
kb	kilo base pairs
ME	Maximum-Entropy
me	Methylation
meDIP	methylated DNA immunoprecipitation

mESC	mouse embryonic stem cell
nm	nanometer
PC	Partial correlation
ph	Phosporylation
PPI	Protein-Protein interaction
ROC	receiver operating characteristic curve
SENet	Sparse Elastic Net
SGL	Sparse Graphical Lasso
SPCN	Sparse Partial Correlation Networks
TF	Transcription factor
TSS	transcriptional start site
ub	Ubiquitinylation

Abstract

Chromatin plays an essential role in transcriptional regulation and in defining cellular identity. Histones, which are the building blocks of chromatin, can be chemically modified with a diverse set of histone modifications. The histone modifications are placed, read or erased by proteins, called chromatin modifiers. Together, chromatin modifiers and histone modifications are components of a chromatin-signaling network involved in transcription and its regulation. The interactions between chromatin modifiers and histone modifications are often unknown, are based on the analysis of few genes or are studied *in vitro*. Further, the functional impact of each chromatin modifier or histone modifications on the whole chromatin signaling network are poorly understood.

With the present thesis, we aim at improving our understanding of the interactions between chromatin modifiers and histone modifications and their function on a genome-wide scale. To this end, we apply computational methods to large sets of genome-wide DNA-protein binding data. From this data we reconstruct the interactions between chromatin modifiers and histone modifications leading to a global chromatin signaling network. First, we evaluate different network reconstruction methods that have been previously applied to genome-wide DNA-protein data on simulated and *Drosophila* data. Second, we provide a high-confidence backbone of the chromatin-signaling network at human promoters. We evaluate the detected interactions in the light of literature knowledge and generate novel biological hypotheses for unknown interactions. Finally, we investigate the differences and commonalities between the chromatin-signaling networks at different chromatin environments in mouse. This analysis results in a systems-level view on the different chromatin signaling interactions leading to novel hypotheses on the functional role of chromatin modifiers and histone modifications in defining the chromatin landscape.

Zusammenfassung

Chromatin spielt eine wichtige Rolle in der Transkriptionsregulation und der Definition von Zelltypen. Die Bausteine des Chromatins, die Histonproteine, können chemisch, mit sogenannten Histonmodifikationen, verändert werden. Die Histonmodifikationen selbst werden von sogenannten Chromatin-modifizierenden Proteinen katalysiert, gelesen oder entfernt. Beide Komponenten zusammen ergeben ein komplexes Chromatin-assoziiertes Signalnetzwerk, das maßgeblich am Transkriptionsprozess und seiner Regulierung beteiligt ist. Die spezifischen Interaktionen zwischen den Histonmodifikationen und den Chromatin-modifizierenden Proteinen sind jedoch meist unbekannt oder basieren auf einer gen-spezifischen oder *in vitro* Analyse. Des Weiteren ist meist die funktionale Bedeutung der einzelnen Interaktionen für das gesamte Chromatin-assoziierte Signalnetzwerk unzureichend bekannt.

Die vorliegende Doktorarbeit hat als Zielsetzung die Interaktionen zwischen Chromatin-modifizierenden Proteinen und Histonmodifikationen, sowie deren Funktion genomweit zu charakterisieren. Mit Hilfe großer genomweiter DNA-Protein-Bindungsdatensätzen und computergestützter Methoden lassen sich die Interaktionen zwischen Chromatin-bindenden Proteinen rekonstruieren. Im ersten Teil der Arbeit werden verschiedene Methoden zur Netzwerkrekonstruktion, die für diesen Zweck in früheren Publikationen verwendet wurden, auf simulierten Daten und *Drosophila*-Daten verglichen. Im zweiten Teil der Arbeit werden Interaktionen an humanen Promoteren rekonstruiert. Diese werden mit Hilfe einer Literatursuche evaluiert und dienen als Basis für neue biologische Hypothesen über bisher unbekannte Funktionen der Histonmodifikationen und Chromatin-modifizierenden Proteine. Im letzten Teil der Arbeit werden Gemeinsamkeiten und Unterschiede zwischen den Chromatin-assoziierten Netzwerken, die für unterschiedlichen Chromatinumgebungen rekonstruiert werden, untersucht. Aus dieser system-orientierten Analyse der Interaktionen lassen sich neue Hypothesen über die Funktion der Histonmodifikationen und den Chromatin-modifizierenden Proteinen bei der Definition verschiedener Chromatinumgebungen ableiten.

Curriculum vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, Juni 2015 Juliane Perner