Freie Universität Berlin

Enhancer Prediction Based on Epigenomic Data

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

vorgelegt von

Anna Ramisch

Berlin, 2019

Erstgutachter: Prof. Dr. Martin Vingron
Zweitgutachterin: Prof. Dr. Silke Rickert-Sperling

Tag der Disputation: 6. Juni 2019

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Berlin, den 12.02.2019

Anna Ramisch

Preface

Acknowledgments

So many amazing people have accompanied me during my PhD journey that it suddenly feels like a huge challenge to acknowledge them all in a way that does them justice. First, I would like to thank Martin Vingron for giving me the chance to discover the beautiful world of bioinformatics. I really appreciate the scientific environment he created, and feel that I grew a lot under his guidance, professionally as well as personally. I would like to thank Martina Lorse and Kirsten Kelleher for their help with basically everything that concerns the organization of a successful PhD time. It was also great to be part of the International Max Planck Research School for Computational Biology and Scientific Computing, which provided me with support as well as friends.

I am very grateful to Alena van Bömmel, who introduced me to Martin and took care of me when I came to the institute with the wish to change my field of study. We share the same experiences from the previous job, and it was good to have her as a role model. Apart from that, she is also one of the happiest and welcoming people I know and is part of the reason why the lab is such a great place. Juliane Perner, Matthew Huska and Alessandro Mammana were my personal (Institute-) Musketeers, and among the people that shaped my first years at the institute the most. We always had a great time together, and while they were teaching me a lot about science, they always had my back and an open ear for my problems. Also Marius Tolzmann and Irina Czogiel made my first years at the institute an unforgettable experience, and I'm grateful for all the moments we helped each other laughing away the one or the other work crisis.

I would like to acknowledge all the people who were part of what is probably the office with the highest 'personnel fluctuation' (my office). The list is long, so I will quickly say 'cheers' to all of them for spending a mostly lovely (sometimes sad) office time with me: Barbara Wilhelm, Katharina Imkeller, Stefan Budach, Florian Massip, Emmeke Aarts, Jessica Walde, Franziska Witte, Philip Kleinert, Marie Coutelier, and Aybuge Altay, who will take over the office now.

I want to thank Matthew Huska and Annalisa Marsico for being great collaborators during the first part of ‘the whole enhancer story’. I learned so much during that time, and really enjoyed working with them. On that note, I want to thank the whole ‘**W**orkgroup for **I**ntegrated prediction of **G**ene expression’, but especially Verena Heinrich, Tobias Zehnder, Philipp Benner and Robert Schöpflin. We delighted and exhausted each other with our enhancer passion to a degree that is hard to understand if you have not been part of it! Verena Heinrich has been, and still is, a wonderful collaborator and friend. We almost finished our big enhancer project that glued us together the past two years, and even though I really cannot see the project anymore, I will miss working on it with her. I am also very grateful to Alisa Fuchs, Sebastiaan Meijsing and my first Bachelor student Yannek Nowatzky, with whom I collaborated the past years on yet another, very interesting, enhancer story. Especially Alisa helped me a lot expanding my biological understanding and, to my delight, shares my enthusiasm for swing dancing.

I want to thank my non-enhancer-related collaborators Matthias Lienhard, Florian Massip and Peter Arndt for the many interesting discussions and for enriching my time in the lab. Of course, Peter deserves extra credit for being the master of coffee and my steady friendly office neighbour. I would like to thank Edgar Steiger for being my friend and math buddy. A coincidence brought us to the same lab after university, and I am really glad that I had him here at my side as my personal consultant.

I am very grateful to my friends Ulli, Anne, Christin, Stephan and Sven, who had nothing to do with work, but had and have everything to do with my well-being and happiness. Especially Ulli, who lived with me for most of my time in Berlin, witnessed many ups and downs during my PhD journey without complaining and is always, and I mean always always, there for me. I would like to thank my partner Guillaume Andrey, who is my solid rock, and helped me a lot during the last months of thesis writing. He is a great source of inspiration and increases my passion for science every day. Finally, I want to thank my family, especially my parents, for their unconditional love and support that enabled everything I achieved so far in life.

Publications

The enhancer prediction method presented in Chapters 4 and 5, as well as its incorporation into a framework for detecting condition-specific regulatory units described in Chapter 6, is based on a collaborational effort with Verena Heinrich, and is made available as a preprint article (Ramisch *et al.*, 2018). The idea for the project started during many previous discussions within the 'workgroup for integrated prediction of gene expression' that several colleagues and I initiated. From these joint discussions my first enhancer-related project started, together with Matthew Huska, Annalisa Marsico and Martin Vingron, in which we predicted enhancers based on a semi-supervised approach and a small set of high-confidence training enhancers. The project resulted in a conference paper (German Bioinformatics Conference 2016) that Matthew Huska and I co-first authored, and which is available at PeerJ Preprints (Huska *et al.*, 2016). The corresponding results are not part of this thesis.

Contents

1 Introduction

An enhancer is a regulatory region which can boost or activate the transcriptional expression of a target gene (Banerji *et al.*, 1981). The target gene can be several kilobases away from the enhancer, upstream or downstream, and is not necessarily the nearest gene to the enhancer. Moreover, a gene can be regulated by several enhancers at the same time (Pennacchio *et al.*, 2013) and one enhancer can regulate several genes (Levine, 2010). In contrast to promoters, the location of most enhancers is still unknown. The majority is located in the non-coding part of the genome, which represents $\sim 98\%$ of the total size, and therefore constitutes a huge search space. Furthermore, most enhancers show dynamically changing activity levels between conditions such as different cell types or time-points. Hence, enhancers that are active in a certain condition can be switched off in another, adding an additional layer of complexity to the task of locating them in the genome (Pennacchio *et al.*, 2013). However, the functional annotation of enhancers and their corresponding target genes is crucial to understand underlying mechanisms of gene expression regulation. In addition to that, enhancers were shown to play a key role in the pathogenesis of many diseases, such as diabetes or certain types of cancer (Wang *et al.*, 2018). Many disease-associated genetic variants are enriched in regions which overlap with known enhancers or show enhancer marks. A subsequent enhancer malfunction can, for example, result in a misregulation of oncogenes (Sur and Taipale, 2016).

While individual enhancers have been validated using reporter assays, it is too time-consuming and expensive to functionally test all enhancer candidates in a corresponding genome-wide approach. However, advances in high-throughput sequencing lead to the exploitation of several genome-wide measurable enhancer properties to serve as a substitute for enhancer activity. Yet, pinpointing an active enhancer based on such data sets, even at a narrowly defined locus, can quickly be too complex to do it manually (see for example Figure 1.1). The more feasible approach is to integrate data representing enhancer properties as features into computational methods for enhancer prediction.

The majority of computational approaches are either unsupervised or supervised. The unsupervised methods search for patterns in the given epigenomic data which subsequently can be used to characterize certain types of genomic elements. No prior knowledge is needed to apply these methods to new data, but the user has to decide which of the discovered patterns or combinations of features 'best' represent the genomic elements of interest, for example active enhancers. Prominent examples of unsupervised genome segmentation methods used for enhancer prediction are ChromHMM (Ernst and Kellis, 2012, 2017) and EpigSeq (Mammana and Chung, 2015), which are based on hidden Markov models, and the dynamic Bayesian network method Segway (Hoffman *et al.*, 2012).

Supervised prediction methods rely on a gold-standard set of known enhancer and non-enhancer regions in the cell type of interest, on which enhancer-associated features distinguishing the one group from the other can be learned. Since often validated enhancers are rare, enhancer properties enter the prediction task through the choice of the feature set as well as through the criteria used to define training enhancers. Both choices should be, at best, independent of each other to avoid circular reasoning. Examples of supervised methods for enhancer prediction are the neural network based approach CSI-ANN (Firpi *et al.*, 2010), or RFECS (Rajagopal *et al.*, 2013) and REPTILE (He *et al.*, 2017) which use random forest classifiers (see Whitaker *et al.* (2015) and Lim *et al.* (2018) for reviews on enhancer prediction methods).

Since one of the challenges in the supervised setting is the construction of a condition-specific training set, the transferability of an already established classifier to other conditions without available gold-standard is crucial. However, only few available prediction tools offer such pre-trained classifiers. Apart from that, the comparison of multiple samples across conditions is often not

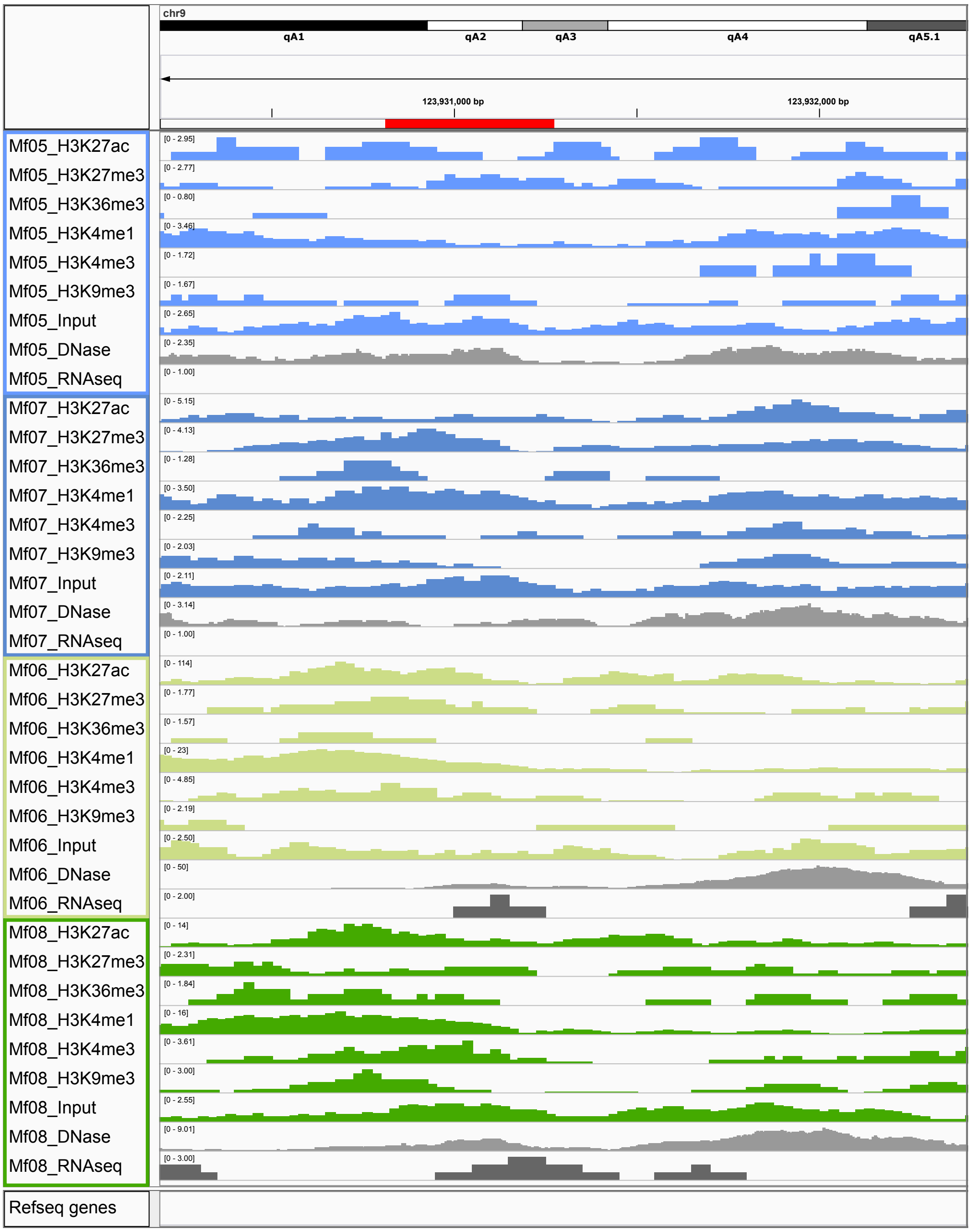


Figure 1.1: Genome browser snapshot. Synovial fibroblast data from two healthy mice ('Mf05', 'Mf07', in blue) and two mice affected with rheumatoid arthritis ('Mf06', 'Mf08', in green) at a putative enhancer (red), which we predicted to be active in the diseased but not in the healthy state. More details on the data can be found in Sections 4.2 and 6.3.

integrated into the available enhancer prediction tools. In fact, the identification and assignment of condition-specific enhancers has to be done by the user in a post-processing step. Another challenging task related to enhancer prediction is the identification of the corresponding target gene. Ernst *et al.* (2011) matched enhancers with target promoters within a defined distance based on the correlation of epigenomic signals across conditions. However, by using a distance-oriented approach they do not take into account the recently revealed partitioning of the genome into domains of preferential chromatin interaction which is thought to limit enhancer activity (Rowley and Corces, 2018). In another approach from Corradin *et al.* (2014), the distance-based criterion for possible enhancer-gene interactions is complemented by binding sites of known insulator elements ($CTCF$) which are involved in the formation of some but not all domain boundaries. Also the motif orientation was not taken into account which is often crucial for the domain formation (Rowley and Corces, 2018).

In this work, we remedy the described problems and present a novel supervised classifier to predict enhancers genome-wide, which can be applied across different conditions without re-training. Furthermore, we integrate our pre-trained classifier into a comprehensive framework to assign enhancer-gene pairs or 'regulatory units' in a condition-specific manner. Our classifier is based on two random forests. It splits the task of distinguishing enhancers from the rest of the genome into two individual ones, focusing especially on the difference between enhancers and promoters. For feature and training set construction, we make use of the widely accepted concept that the condition-specific activity of an enhancer can be characterized by an accessible region flanked by nucleosomes which carry specific histone modification (HM) patterns (Heintzman *et al.*, 2009; Rada-Iglesias *et al.*, 2011). Our classifier is based on six core histone modifications, which are available for many different cell types and species (Bernstein *et al.*, 2010; Stunnenberg *et al.*, 2016) and therefore guarantee a broad applicability. Furthermore, the design of the feature set takes the particular shape of the HM distribution at active enhancers into account by including several bins around the center of the enhancer. For the definition of our training enhancers, we also include bidirectional transcription as an additional mark of active enhancers as well as chromatin accessibility. Thousands of condition-specific bidirectionally transcribed enhancers have been experimentally identified and collected in the FANTOM5 database (Andersson *et al.*, 2014), such that we were able to construct training sets in different cell types and species.

In a collaborative effort with Verena Heinrich, our pre-trained classifier is integrated into a comprehensive framework to predict condition-specific regulatory units (Ramisch *et al.*, 2018). The classifier is applied to histone modification data from several conditions, and differentially active enhancer regions are identified using a non-parametric permutation test directly on the predicted enhancer probabilities. Subsequently, the differential enhancers are assigned to their corresponding conditions and matched to putative target genes based on a high correlation between their probabilities and gene expression values. Here, the search space for each enhancer is restricted to its topologically associating domain (TAD) to take into account prior knowledge about domains of preferential chromatin interaction. The final outcome is a manageable candidate list of condition-specific enhancer-gene pairs that can be used for further analyses.

Outline of this work

In Chapter 2, we give a summary of the biological background of this work. We introduce a collection of experimental techniques and the corresponding measurable enhancer properties which we exploit for the design and training of our prediction method as well as the prediction of target genes. In Chapter 3, we give a general overview of machine learning and the two main types of algorithms: unsupervised and supervised methods. Furthermore, we introduce several performance measures which we used for validation or method optimization. Chapter 4 presents the construction, and Chapter 5 the validation and comparison of our enhancer prediction method. We cover both prediction within a cell type as well as across different conditions, and finally offer a pre-trained classifier. In Chapter 6, we introduce our framework to predict condition-specific regulatory units and apply it to a complex disease model.

2 Biological Background

2.1 Regulation of gene expression

Gene expression is the process in which the encoded information from a gene is first transcribed from DNA into RNA and then translated to a functional gene product, e.g. a protein. The identity and function of a cell depends on the regulation of gene expression, which comprises control mechanisms to adjust the rate and time of transcription and translation for each individual gene. During development, the regulation of gene expression is critical to produce the diversity of cell types (Alberts *et al.*, 2002). Moreover, even cells of the same type can show differential patterns of gene expression, i.e., different concentration levels of the same protein or functional RNA, as a response to environmental stressors.

The primary control point for the regulation of gene expression is thought to be the initiation of transcription. Transcription starts at the **transcription start site** (TSS) which is located immediately at the $5'$ end of a gene. The TSS is part of the **promoter**, a short asymmetric sequence which is able to assemble the transcriptional machinery including RNA polymerase II and a set of proteins called **general transcription factors**. The promoter of a gene dictates the accurate position of initiation as well as the direction of transcription. The rate of transcription initiation is flexible and can be influenced by many factors, for example by the speed of the transcriptional machinery assembly or the presence of **gene regulatory proteins** including transcriptional activators. The set of regulatory proteins can differ for each gene, as well as the location of their binding sites which can be close to the promoter or further away. In fact, these regulatory sequences called **enhancers** are able to affect the RNA polymerase activity of a gene from distances of hundreds of thousand of base pairs away and therefore add an additional layer of regulatory potential (Alberts *et al.*, 2002).

2.2 Enhancer definition and short history

Enhancers are genomic regions of up to several hundred base pairs (bp) in length that are able to boost the transcription of a target gene independently of their orientation or distance to the target promoter (Banerji *et al.*, 1981). The first enhancer, a 72 bp repeated sequence element, was found in a small DNA virus more than 30 years ago, more or less by coincidence by Banerji *et al.* (1981). The original aim of the corresponding study was to analyze the effects of alterations in eukaryotic gene promoters on the gene expression to identify putative regulatory elements. In the experiment which finally led to the enhancer discovery, a 5000 bp fragment of rabbit DNA containing the β-globin gene was cloned into two vectors, one including an SV40 DNA segment (intended for DNA amplification). When transfecting into Hela cells, only the SV40-containing plasmid produced high levels of β-globin. Follow-up experiments using different constructs showed that the boost of globin expression was highly reproducible, irrespective of the orientation of the insert at the cloning site, and at distances of more than 1000 bp (1400 bp upstream or 3300 bp downstream) from the TSS of the gene. These functional properties observed in 1981 are still the foundation of the currently accepted enhancer definition (see e.g. Pennacchio *et al.* (2013) for an enhancer review).

Shortly after, the first non-viral enhancer was discovered in the immunoglobulin heavy chain (IgH) gene locus, which was the first cell type-specific genetic element of that kind (Banerji *et al.*, 1983; Gillies *et al.*, 1983). The development of so-called enhancer traps, which can determine the enhancer activity of randomly-selected regions, provided a sophisticated and systematic approach to define enhancer regions (Weber *et al.*, 1984; Hamada, 1986; Bellen *et al.*, 1989). In the following years, various properties of enhancers were observed and with an increasing number of novel methods these properties could be studied in more detail.

2.3 Experimental techniques commonly used to study gene regulation

2.3.1 ChIP-seq

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) was first described by Barski *et al.* (2007) and allows the genome-wide identification of binding sites for TFs, histones and other proteins *in vivo*.

DNA-bound proteins are cross-linked to the chromatin using formaldehyde thereby producing covalent bonds between them. Then, the DNA is fragmented using ultrasounds (sonication), and fragments which are linked to the protein of interest are subsequently isolated using an antibody recognizing the specific protein or protein modification. The purified DNA-protein complexes are then reverse cross-linked and the DNA is prepared for deep-sequencing. The pulled-down DNA fragments are sequenced using next-generation sequencing (NGS) and mapped back to the genome to identify the genome-wide protein binding sites.

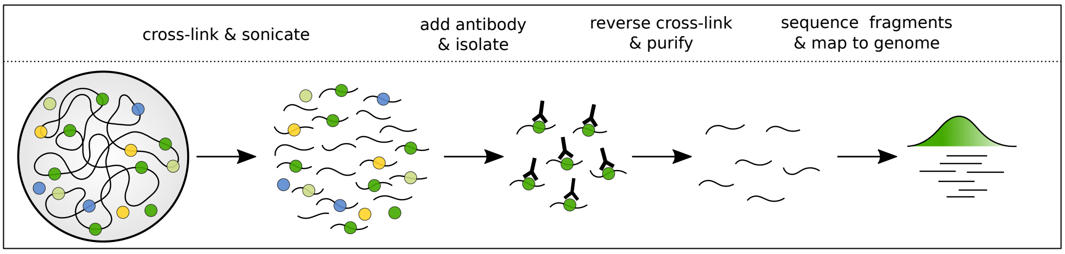

cross-link & sonicate
add antibody & isolate
reverse cross-link & purify
sequence fragments & map to genome

Figure 2.1: Overview ChIP-seq protocol. Chromatin in the nuclei is cross-linked and fragmented. DNA fragments bound to the proteins or histone modification of interest (depicted as green balls) are enriched in the immunoprecipitation step by the specific antibody. Finally, the isolated fragments are reverse cross-linked, purified and sequenced. Inspired by Barski *et al.* (2007).

2.3.2 RNA-seq

RNA sequencing (RNA-seq) is used to measure the transcriptome of a cell, which comprises the complete set of transcripts and their isoforms, as well as their quantity at a specific condition or developmental stage (Lister *et al.*, 2008; Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008).

The workflow of an RNA-seq experiment starts with isolating the RNA of interest from a given organism and cell type. The isolated RNA molecules are reverse transcribed into cDNA and fragmented (or in an inverted order, first fragmented and then reverse transcribed). The cDNA fragments of the desired size are ligated with adapters to create the cDNA input library. In a last step, the library is amplified and sequenced using high-throughput sequencing technology.

Most differences in RNA-seq technologies can be found in the library preparation step. Some approaches of adding sequencing adapters lack strand specificity, while others assure that this important information is kept for amplification and sequencing. An overview of strand-specific RNA-seq methods can be found in Levin *et al.* (2010) or Hrdlickova *et al.* (2017).

2.3.3 CAGE

Cap analysis gene expression (CAGE) published by Shiraki *et al.* (2003) is a high-throughput gene expression technique to identify transcriptional starting points genome-wide by identifying capped 5' ends of coding and noncoding mRNA. The standard protocol was updated by Takahashi *et al.* (2012) and can be summarized as follows.

First, cDNA is synthesized from an mRNA population of interest and full-length cDNA are captured by linking/trapping a biotin residue at the 5' cap structure (cap trapper method). Incompletely synthesized cDNAs are eliminated in this step. Then, 27 nucleotides are cleaved inside the 5' end of the cDNAs by digestion of the restriction-modification enzyme EcoP15I. The resulting CAGE tags are amplified in a PCR step and sequenced.

2.3.4 DNase-seq

DNase I hypersensitive sites sequencing (DNase-seq) is a method to identify DNase I hypersensitive sites (DHSs) genome-wide as a proxy for open chromatin regions (Crawford *et al.*, 2006; Boyle *et al.*, 2008). First, nuclei of a selected cell population are isolated and treated with a DNase I concentration to release short DNA fragments. Then, the DNase-digested fragments are isolated and ligated to sequencing adapters for NGS library preparation.

2.3.5 ATAC-seq

Assay for transposase accessible chromatin with high-throughput sequencing (ATAC-seq) is a relatively new method from Buenrostro *et al.* (2015) that can be used as an alternative to DNase-seq to measure whole-genome chromatin accessibility. DNA samples of interest are exposed to the mutated hyperactive transposase Tn5 which preferentially cuts at open regions and simultaneously ligates sequencing adapters. The resulting transposed DNA fragments are then isolated, amplified by PCR and sequenced.

2.4 Enhancer properties

After the discovery of the first enhancer and the establishment of a functional enhancer definition (see Section 2.2), various properties of enhancers were observed that are closely intertwined with each other. Many experimental methods, some of them reviewed in Section 2.3, and novel computational methods were developed to study these properties. Below, we first give a short summary of several observed enhancer features and then discuss each individual feature in more detail in the following sections.

2.4.1 Overview of properties

Active enhancers act as binding platforms and are accessible to a combination of various transcription factors (TFs) (Long *et al.*, 2016). The flanking nucleosomes were observed to carry specific histone modifications (HMs) such as H3K4me1 or H3K27ac (Rada-Iglesias *et al.*, 2011). Active enhancers are often hypomethylated, i.e. show low methylation levels (Schmidl *et al.*, 2009), and frequently produce short RNA transcripts in a bidirectional fashion (Melgar *et al.*, 2011). A particularly active type of enhancers are super-enhancers, which correspond to large genomic segments displaying abundant characteristic enhancer features (Hnisz *et al.*, 2013; Whyte *et al.*, 2013).

The activity of enhancers is mostly condition-specific, and as such enhancers display differential activities between different conditions. In accordance with activity, also the chromatin accessibility as well as the characteristic HM pattern of an enhancer change dynamically between conditions (Heintzman *et al.*, 2009). A subset of enhancers overlap with differentially methylated regions (DMRs) and can display hypermethylated or hypomethylated CpG dinucleotides depending on their activity (Schmidl *et al.*, 2009).

Enhancers can regulate genes independently of their orientation and over very large genomic distances. However, the capacity of enhancers to control gene transcription is thought to be restricted to genes located in the same domain of preferential chromatin interaction (Rowley and Corces, 2018). Finally, some enhancers, but not all, are highly conserved between species (e.g. Blow *et al.* (2010)). A simplified representation of several enhancer features and their dynamics between conditions can be found in Figure 2.2.

2.4.2 Cell type-specific histone modifications

Nucleosomes form the fundamental repeating units of eukaryotic chromatin. Each nucleosome consists of approximately 146 bp of DNA wrapped around a histone octamer consisting of two copies each of the core histones H2A, H2B, H3, and H4. Each histone protein has an unstructured tail extended from the globular domain which can be post-translationally modified. Post-translational histone tail modifications such as acetylation and methylation have been observed and documented for many years and were thought to affect chromatin structure by influencing histone-DNA or histone-histone contacts (van Holde, 1989). The observation of distinct, condition-dependent patterns of histone modifications (HMs) endorsed the idea that histones are not only, as

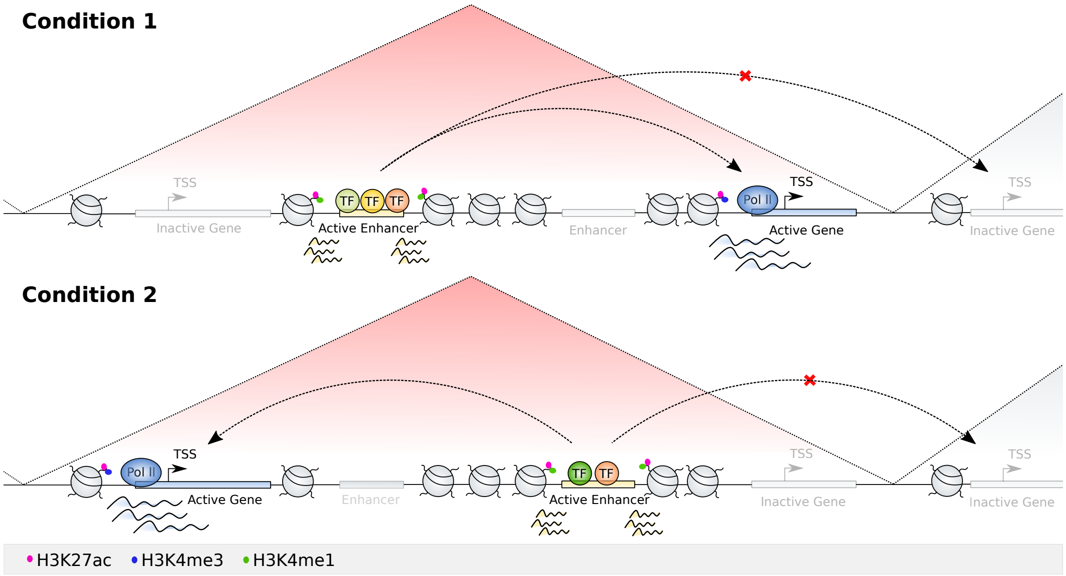


Figure 2.2: Overview of enhancer properties. Genomic landscape for two different conditions. Active enhancers (yellow) are bound by a combination of transcription factors, produce bidirectional short transcripts and are flanked by nucleosomes carrying H3K27ac and H3K4me1. The active target genes (light blue) are either upstream (condition 1) or downstream (condition 2) of the regulating enhancer (indicated with arrows) and are located in the same domain of preferential chromatin interaction (red triangle). Histone modifications, transcript production and transcription factor binding is changing dynamically between the conditions according to the changes in enhancer activity (inactive enhancers in light grey).

originally thought, static structural elements but dynamic components linked to gene regulation. Furthermore, it was proposed that HMs may act together in a complex sequential or combinatorial manner to encode a 'language', the so-called **histone code**, that can be read by other proteins (Strahl and Allis, 2000).

The specific functions of most of the HMs, however, were still unknown at that point, especially due to the lack of information regarding the responsible histone-modifying enzymes. Histone acetylation (of histone H4) was already very early linked to transcriptionally active genes (Hebbes *et al.*, 1988), and the identification of the first histone acetyltransferase (HAT) initiated intensive functional studies of histone acetylation in connection with transcriptional regulation (Brownell *et al.*, 1996). The same holds true for histone methylation, where the finding of H3-specific methyltransferases pioneered the research to link methylation of certain histone tails to gene activity (Chen *et al.*, 1999; Rea *et al.*, 2000).

Locus-centric studies served as a first test bed for hypotheses about the properties of the histone code. The β-globin locus, for example, was studied extensively (Litt *et al.*, 2001; Agalioti *et al.*, 2002; Bulger *et al.*, 2003), as well as other well characterized developmental genes (Schneider *et al.*, 2004). With the arrival of new technologies like ChIP-seq (see Section 2.3.1 for more details), larger regions could be studied with increasing resolutions until genome-wide analyses became standard procedure also for higher eukaryotes.

Scientists observed high levels of histone H3 acetylation and H3K4me3 at the **promoter regions** of active genes (Schübeler *et al.*, 2004; Kim *et al.*, 2005; Liang *et al.*, 2004) which could be correlated with the level of chromatin accessibility and gene expression (Roh *et al.*, 2005; Bernstein *et al.*, 2005). In embryonic stem cells (ESCs), H3K4me3 and the repressive mark H3K27me3 were found to colocalize at promoters of developmental genes to form a so-called 'bivalent' configuration which is suggested to keep the genes ready or 'poised' for a rapid activation or repression (Roh *et al.*, 2006; Boyer *et al.*, 2006; Bernstein *et al.*, 2006; Barski *et al.*, 2007).

Many of the HMs enriched at promoters were also detected in intergenic or transcribed regions, for example H3 acetylation which could be observed at known functional **enhancers** (Roh *et al.*, 2005). The first mark distinguishing enhancers from promoters was H3K4me1 (Heintzman *et al.*, 2007; Wang *et al.*, 2008). Later, Rada-Iglesias *et al.* (2011) observed a group of inactive enhancers in ESCs, so-called 'poised' enhancers, that could be characterized by an absence of H3K27ac, an enrichment of the active mark H3K4me1 and an enrichment of the repressive mark H3K27me3. This finding refined the role of H3K27ac, together with H3K4me1, as the active enhancer mark. In contrast to promoters, the enhancer-associated HM patterns were largely cell type-specific (Heintzman *et al.*, 2009).

Very recently, a new class of active enhancers has been observed which show a lack of H3K27ac, but are enriched for the globular domain acetylation H3K122ac instead (Pradeepa *et al.*, 2016).

2.4.3 The role of chromatin accessibility in enhancer function

Already before the discovery of the first enhancer in 1981, it was observed that certain regulatory regions are hypersensitive to DNase I digestion, and as such are nucleosome-free (Wu, 1980). The exposed DNA sequences are thus accessible for recognition and binding by TFs. As a result, DNase I hypersensitivity mapping has been extensively used to identify ubiquitous and tissue-specific regulatory regions, including enhancers. However, the first effort of mapping DNaseI hypersensitive sites (DHSs) within the chromatin required a lot of tedious working steps, was quite inaccurate and was limited to single loci. In combination with array technology and later high-throughput sequencing approaches, such as DNase-seq (see Section 2.3.4), genome-wide libraries of DHSs could be generated and many new putative enhancers identified using various peak-calling algorithms (Crawford *et al.*, 2006; Sabo *et al.*, 2006; Boyle *et al.*, 2008; Hesselberth *et al.*, 2009; Bernstein *et al.*, 2010).

In the mean time, also other methods to detect open regions emerged which were used for enhancer identification as well (Fu *et al.*, 2018; Daugherty *et al.*, 2017; Davie *et al.*, 2015), for example formaldehyde-assisted isolation of regulatory elements combined with high-throughput sequencing (*FAIRE-seq*, Giresi *et al.* (2007)) or *ATAC-seq*, which is short for assays for transposase accessible chromatin and high-throughput sequencing (Buenrostro *et al.* (2015), see Section 2.3.5).

2.4.4 Enhancers as platforms for transcription factor binding

Enhancers can be considered as clusters of short DNA sequences called motifs, that can be specifically recognized by DNA binding TFs. The bound TFs can recruit co-activators or co-repressors, which often lack a sequence-specific DNA-binding competency. These co-factors, in turn, can modify the chromatin nearby the enhancer and subsequently determine its transcriptional activation (Buecker and Wysocka, 2012; Long *et al.*, 2016).

In addition to the combination of several motifs, there are many other parameters which can affect the functional output of an enhancer. These universal principles of motif organization within an enhancer, often referred to as **enhancer grammar**, comprise the number and the affinity of individual binding motifs, the spacing between them, their order and orientation, as well as the local DNA shape (Spitz and Furlong, 2012; Long *et al.*, 2016). How flexible these principles are is still a topic of research, and close inspection of the architecture of individual well-studied enhancers has currently led to three distinct models to describe enhancer activity which are also depicted in Figure 2.3. The **enhanceosome model** relies on a rigid motif composition and grammar, which means that all recruited TFs and their relative positioning are essential to produce a functional outcome (Thanos and Maniatis, 1995; Merika and Thanos, 2001). The **billboard model**, on the other hand, allows for more flexibility. Even though the motif composition is fixed, there are fewer constraints on the relative positioning of the motifs. Also, the binding of (different) subsets of TFs at the enhancer site can be sufficient for its activation (Kulkarni and Arnosti, 2003). The **TF collective model** describes the scenario in which a specific set of TFs binds (as a collective) to a set of enhancers without leading to any obvious shared motif grammar. In this recruitment situation, usually all TFs are necessary for enhancer activity, but the grammar is very flexible since only a (varying) subset of TFs binds directly to the DNA and the rest are recruited through protein-protein interactions (Junion *et al.*, 2012). In fact, most enhancer architectures may constitute a mixture of the three models where some TFs rely on a rigid motif grammar but not others (Ng *et al.*, 2014).

In summary, this suggests that a given enhancer activity or output can be generated by multiple motif compositions and organizations, and in reverse, a given combination of recruited TFs can generate multiple functional enhancer outputs depending on their relative positioning (Spitz and Furlong, 2012).

This lack of generalizable motif rules at enhancers makes it difficult to uncover the often cell type-specific TF combinations necessary for activation. However, existing knowledge of key signaling pathways coupled with high-throughput technology made it possible to examine well-studied cell types for enhancer-

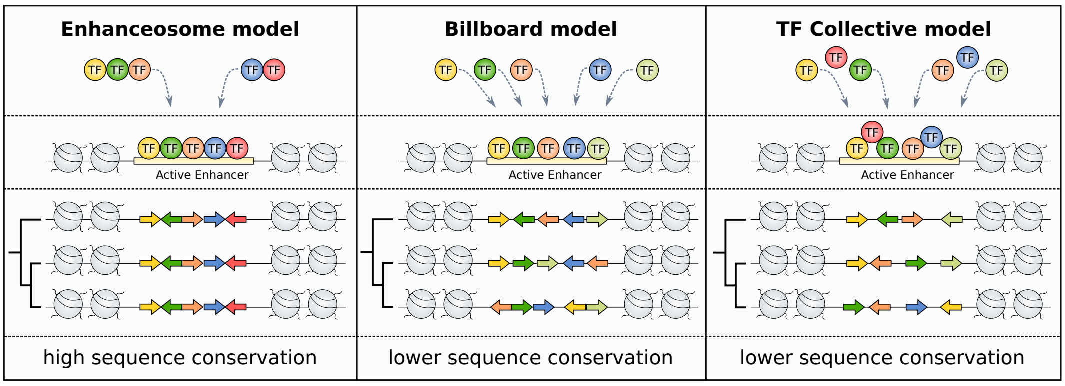

Enhanceosome model
Billboard model
TF Collective model
TF
Active Enhancer
high sequence conservation
lower sequence conservation
lower sequence conservation

Figure 2.3: Motif architecture models for enhancers. In the enhanceosome model, TFs bind by direct cooperativity, i.e., a direct protein-protein interaction, directly to the DNA. The motif organization is very stringent and the evolutionary conservation high. In the billboard model, a fixed type and number of TFs bind by indirect cooperativity directly to the DNA, while their motif organization is flexible. In the TF collective model, TFs bind by indirect cooperativity directly and indirectly to the DNA, resulting in varying motifs being present. For both, the billboard and the TF collective model, the evolutionary conservation is low. Inspired by Long *et al.* (2016)

associated TFs. In mouse embryonic stem cells (mESCs), for example, ChIP-seq experiments of sequence-specific TFs revealed that regions enriched in known pluripotency factors showed enhancer activity (Chen *et al.*, 2008).
Moreover, since activating co-factors are recruited by single TFs irrespective of cell type, they are often preferred for genome-wide enhancer identification, especially when prior knowledge of the underlying regulatory network is sparse. The acetyltransferase and transcriptional co-activator *p300* is one of few known enhancer-associated factors. It was found to be present at both active and poised enhancers (Buecker and Wysocka, 2012), and is widely used to predict the location and cell type-specific activity of enhancers (e.g. Visel *et al.* (2009)). The Ada-Two-A-Containing complex is another co-activator with a known enhancer association. It is a multiprotein co-activator that contains a catalytic histone acetyltransferase (HAT) module and can be found at promoters as well as enhancers. Its enhancer binding is cell type-specific and interestingly defines a class of enhancers not bound by *p300* (Krebs *et al.*, 2011).

2.4.5 Functional rather than sequence conservation

Nonfunctional DNA collects mutations much faster than functional DNA, since deleterious mutations are generally eliminated by natural selection while mutations which have no phenotypic or only slightly deleterious effects can be randomly fixed in the population. As a consequence, most of the highly conserved sequences, i.e., sequences which are maintained over evolution and do not accumulate mutations, are very likely functional (Charlesworth, 2012).

Following this logic, one of the first efficient computational approaches to predict enhancers was based on inter-species comparisons to pinpoint conserved regions. Nobrega *et al.* (2003), for example, compared two gene deserts surrounding the human *DACH* gene across multiple species to identify evolutionary conserved regions corresponding to putative *DACH* enhancers. Applying new computational methods they were able to find several conserved sequences which they also tested for *in vivo* enhancer activity. The increasing number of available whole genome sequences paved the way for advanced comparative genomic tools to predict enhancers with a high specificity enabling an *in vivo* characterization and assembly of the predicted regions (Siepel *et al.*, 2005; Prabhakar *et al.*, 2006). In fact, the VISTA database was the first public resource which provided access to evolutionary conserved noncoding human sequences tested for enhancer activity using transgenic mouse assays (Visel *et al.*, 2007). Many noncoding sequences, especially in and around genes associated with vertebrate development, were found to be enriched for enhancers (Woolfe *et al.*, 2004; Pennacchio *et al.*, 2006; Prabhakar *et al.*, 2006).

On the other hand, there is a substantial fraction of identified enhancers with only a modest or no conservation at all (Schmidt *et al.*, 2010; Blow *et al.*, 2010; May *et al.*, 2012). The level of conservation was found to vary depending on the tissue-specificity of an enhancer (Blow *et al.*, 2010), but also depending on its activity status. Poised developmental enhancers, for example, exhibited an overall low conservation (Rada-Iglesias *et al.*, 2011). Moreover, the evolutionary sequence conservation of enhancers is closely connected to its motif architecture (see also Figure 2.3).

Recently, Arnold *et al.* (2014) found that a large fraction of enhancers show deeply conserved activity as a result of selection and transcription factor binding site (TFBS) turnover, even between orthologous regions of evolutionary close species. This pinpoints towards a **conservation of function** of enhancers rather than conservation of their underlying DNA sequence.

2.4.6 Bidirectional transcription

During investigations of the transcriptional landscape of higher eukaryotes it became clear that not only protein coding sequences are transcribed, but a large portion of the non coding genome as well. (Carninci *et al.*, 2005; Cheng *et al.*, 2005; Johnson *et al.*, 2005; The ENCODE Project Consortium, 2007; Kapranov *et al.*, 2007). These noncoding transcripts reside in intergenic regions as well as in introns of known genes, and their discovery raised a lot of questions and controversies about the origin and possible functionality of the 'dark matter' or 'pervasive' transcription (Johnson *et al.*, 2005; Struhl, 2007; Ponjavic *et al.*, 2007; van Bakel *et al.*, 2010; Clark *et al.*, 2011).
Genome-wide studies of RNA polymerase II occupancy revealed widespread transcription at enhancers and showed that, in fact, enhancers build the major group of noncoding regions undergoing transcription (Kim *et al.*, 2010; De Santa *et al.*, 2010; Koch *et al.*, 2011). Most of the enhancer transcripts or **eRNAs** are relatively short (0.5-2 kb), nonpolyadenylated and bidirectional, but a subset of enhancers was also found to generate undirectional, rather long ($>$ 3-4 kb) and polyadenylated eRNAs (Kim *et al.*, 2010; Wang *et al.*, 2011; Natoli and Andrau, 2012). Nevertheless, bidirectional transcription became a hallmark of enhancers and was widely used to (computationally) predict enhancers in a genome-wide manner (Melgar *et al.*, 2011). The **FANTOM5 consortium** (Functional Annotation of the Mouse/Mammalian Genome) provides an atlas of *in vivo* transcribed enhancers (and promoters) across multiple tissues and cell types in human and mouse. Based on CAGE data (see 2.3.3 for more details) they mapped bidirectional transcription and identified the corresponding TSSs in the genome (Andersson *et al.*, 2014). In these studies, it was also shown that the amount of produced bidirectional eRNA correlates well with enhancer activity and therefore is a good indicator for cell type-

specificity.

In addition, the findings suggest that the transcription or the produced eRNAs themselves might have a mechanistic importance for the activity of the corresponding enhancer. And indeed, experimental evidence has been found for the functional importance of several candidate eRNAs. In some cases, the eRNA molecules seem to be necessary for enhancer activity, since their degradation lead to reduced nearby mRNA expression (Lam *et al.*, 2013) or their induction caused an increase in enhancer-promoter looping strength (Li *et al.*, 2013). Furthermore, depletion of eRNAs could be linked to a decrease in transcript and protein levels of nearby genes (Ørom *et al.*, 2010; Mousavi *et al.*, 2013), and eRNAs were shown to promote RNA polymerase II recruitment to certain genomic loci (Johnson *et al.*, 2003; Mousavi *et al.*, 2013).

However, despite (few) examples of functional enhancer transcripts there is a clear reporting bias due to the fact that it is difficult to prove the non-functionality of an eRNA.

It was also recently discussed that bidirectional transcription is not specific to enhancers, but solely a mark for accessible chromatin (Young *et al.*, 2017). As such, it is a feature found at many active enhancers, but also at regions which do not show any enhancer-associated chromatin marks or enhancer activity.

2.4.7 Variable DNA methylation

DNA methylation is an epigenetic modification in which a methyl group is added to the DNA molecule, either to the cytosine or adenine base, without changing the underlying sequence. In eukaryotes, the methylation of cytosine is much more prevalent and was already mentioned in 1975 as a possible regulatory key player in development or X inactivation, making it one of the most studied modifications (Riggs, 1975; Holliday and Pugh, 1975). Cytosine methylation has been found in different sequence contexts, but in mammals it predominantly occurs at CpG dinucleotides (Ziller *et al.*, 2011). **CpG islands** (CGIs) are genomic regions which show a high CG density, a high frequency of CpG sites and are mostly located at promoters of housekeeping and developmentally regulated genes (Smith and Meissner, 2013). Most of the CGIs at promoters are not methylated, but if they are, the methylation is associated with long-term silencing of the genes (Jones, 2012). Enhancers are mostly CpG-poor and many were observed to be neither completely methylated nor unmethylated, hence termed 'low-methylated regions' or LMRs (Stadler *et al.*, 2011). This is a result of averaging over dynamically changing (binary) methylation states of the individual CpGs located in the enhancer (Jones, 2012). Additionally, **differentially methylated regions** (DMRs), which display a varying methylation status across conditions like different cell types or time-points, were found to overlap with enhancer regions (Ziller *et al.*, 2013). It was shown for a subset of DMRs that hypomethylation associates with an increased regulatory activity, as the reverse holds true for hypermethylation (Schmidl *et al.*, 2009).

2.4.8 Enhancer-promoter communication is restricted to topologically associating domains

The genome is divided into different chromosomes, and the chromosomes themselves are partitioned into physically and structurally distinct domains which relate to their position in the nucleus and their local organization in the three dimensional space, respectively. These domains are also associated to their level of activity. The non-random organization of the genome in the (mammalian) cell nucleus has been studied for more than three decades (see Cremer and Cremer (2001) for a comprehensive review), and has been shown to constitute an important layer of gene regulation since the direct physical environment of a gene has the potential to affect its expression (Bell *et al.*, 2001). Genes closer to the nuclear periphery, for example, are more often found in a repressed state than those located further away. Moreover, repositioning of genes to the periphery leads to reduction in gene expression (Finlan *et al.*, 2008).

At the megabase-scale level of chromosomes, chromatin is organized in insulated domains. Scientist have postulated the existence of insulator elements or barriers, which can block the spreading of active and inactive chromatin states (Sun and Elgin, 1999; Bell *et al.*, 2001). **Insulators** are of special interest in the context of enhancers, since they are neutral barriers able to block enhancer function when positioned between the enhancer and its target promoter, but have no functional consequence when positioned outside of the region which lies between the two regulatory elements (Bell *et al.*, 2001).

It is now thought that insulators mostly function in vertebrate genomes by creating domains of preferential chromatin interactions (Rowley and Corces, 2018). These domains were revealed using genome-wide mapping of chromatin interactions based on chromosome conformation techniques (like Hi-C and 5C) and are referred to as **topologically associating domains** (TADs) (Lieberman-Aiden *et al.*, 2009; Nora *et al.*, 2012; Dixon *et al.*, 2012). The resulting partitioning of the genome corresponds to 'regulatory neighborhoods' limiting enhancer activity to genes located in the same TAD. It was also found, that the TAD structure is (to a certain degree) invariant between cell types and even conserved between species (Dixon *et al.*, 2012).

Apart from being insulated regulatory units of the genome, TADs are molds for enhancer-promoter communication, enabling the normal regulation via the formation of chromatin loops. The looping model is one of the possible explanations hypothesizing that the communication of enhancers and promoters was driven by the interaction of proteins bound to both regulatory elements resulting in a loop of the intermediate DNA (Ptashne, 1986). An extensive set of evidence has endorsed this model for over a decade now. One of the first evidence thereof was the β-globin gene (Tolhuis *et al.*, 2002). Here, using chromosome conformation capture (3C)-based assays allowed to capture spatial proximity between specific pairs of loci in the nucleus. Enhancer-promoter loops were also observed at thousands of loci in the following years (Javierre *et al.*, 2016).

2.4.9 Super-enhancers

Recently, it was observed in several mammalian cell lines that a subset of enhancers cluster together in large domains, called **super-enhancers**, which show high levels of H3K27ac, enhancer-associated TF binding as well as binding of the Mediator coactivator (*Med1*), and primarily control cell identity genes (Hnisz *et al.*, 2013; Whyte *et al.*, 2013).

Whyte *et al.* (2013) defined super-enhancers in mESC based on *Med1* ChIP-seq occupancy. In their analysis, enhancers were called based on TF enrichment and stitched together within a range of 12.5 kb. Then, based on the distribution of the coactivator *Med1* over all enhancers a cutoff was chosen, to define super-enhancers. In mESC, *Med1* was found to be the 'optimal' decision factor, since it showed the clearest transition between super-enhancers and other 'typical' enhancers in comparison to DNase I, H3K27ac or H3K4me1. The mESC super-enhancers have a larger size than typical enhancers (median distance of $\sim 8,6$ kb vs. ~ 700 bp), show high ChIP-seq occupancy of *OCT4*, *SOX2*, *NANOG*, *KLF4* and *ESRRB*, and regulate genes which are know to control the pluripotent state of the cell.

3 Machine Learning Background

3.1 Introduction

The term **machine learning** was first introduced by Samuel (1959) and evolved out of the field of artificial intelligence. The focus of machine learning is, quite literally, the creation of 'machines' (e.g. in the form of algorithms) that 'learn' from a given large data set by extracting meaningful information and gaining experience to increase their performance in the future (see Chapter 1 of Izenman (2008)). Since there is a need for understanding and learning from large data sets in several applications, the techniques employed in machine learning are also influenced by many different disciplines such as statistics, pattern recognition or artificial intelligence, to name only a few examples. As a consequence, there are several possibilities how to categorize the multitude of developed machine learning approaches. In this work, we will distinguish different approaches according to their learning strategies into two main groups: supervised and unsupervised learning. The goal in **supervised learning** is to to find a function that approximates a provided correct output variable from a given set of input variables. This approach is comparable to learning with a teacher who knows the correct answers and can therefore help to improve results progressively (by repetition) until a certain performance is reached. The input and output variables can be of categorical or continuous nature whereby depending on the nature of the output variable, supervised learning splits into regression (continuous) or classification (categorical). In **unsupervised learning** the algorithm is also provided with (categorical or continuous) input variables, but not with a correct output variable. Thus, unsupervised learning pursues a different goal in that it explores properties of the input. It is comparable to the scenario in which no teacher is available to provide any corrections. The different assumptions in supervised and unsupervised learning methods demonstrate that the choice of method type is motivated by data availability. But this is just one aspect in a complex set of decisions leading to a successfully solved learning problem. In Chapter 1 of Duda *et al.* (2001) this is described as a **design cycle**, a repetition of data collection, feature choice, model choice, training, and evaluation steps.

Data collection is the first part of the design. A lot of data is usually needed to achieve a good performance, but its collection is often costly and time consuming. The **choice of features** specifies the part of the information decoded in the data which the learning algorithm will have access to, which usually requires some prior knowledge about the nature of the data. There are also some desirable properties a good feature set should fulfill: it should be easy to extract, insensitive to noise and also useful for actually making the discrimination of interest. The **model choice** is more than just deciding if the approach is unsupervised or supervised. Within each of this learning types there are several methods to choose from, where each of them comes with advantages and disadvantages, strength and weaknesses. Hence, this step is also very dependent on prior knowledge of the collected data and chosen features.

Training is the process of inferring knowledge from the collected data and make predictions based on what was learned. There are different ways how to train a model, which are mostly connected to data availability but also to the model choice. The **performance evaluation** of the trained model is important for multiple reasons. It helps with comparing different models (and hence can influence the final model choice) and it can tell us if there is still a lot of room for improvement. Evaluation is also crucial to find a good trade-off between model complexity and the ability of a model to generalize. The more complex a model gets, the better it might be able to explain the training data, but a likely consequence is that it performs poorly on new unseen data. This phenomenon is called overfitting and stands in contrast to the main goal of machine learning which is to make reliable predictions on future data.

3.2 Mathematical setting

Let X and Y be random variables which can be represented by a joint distribution $\mathbb{P}(X, Y)$. Here, $X = (X_1, \ldots, X_p) \in \mathcal{X}$ denotes the set of input variables of interest, which can be categorical or continuous. Each variable describes a certain feature of the input data set and is therefore also called input **feature**. The number of features, p, describes the dimensionality the input data lives in. Y is the output or response variable and can also be of continuous or categorical nature.

The aim of **unsupervised learning** is to infer the properties of the joint probability density $\mathbb{P}(X)$ from an unlabeled input data $\mathcal{U} = \{x_1, \ldots, x_n\}$, since there are no data realizations of Y available. In **supervised learning**, the goal is to predict outcomes of Y from X by determining the properties of the conditional density of Y given X, $\mathbb{P}(Y|X)$, from the labeled learning set $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. For continuous Y, the prediction task is called **regression**, while for a categorical Y it is called **classification** (see e.g. Hastie *et al.* (2009), Chapters 2 and 14).

In this work, we will concentrate on the **binary classification problem** where Y can take two possible classes, $Y \in \mathcal{Y} = \{0, 1\}$. Hence, our classification task is to learn a function $h_{\mathcal{L}} : \mathcal{X} \to \{0, 1\}$ on the training set $\mathcal{L}$ which predicts the class membership of an input observation $x \in \mathcal{X}$. Since we are focusing on binary classification, we model the conditional class probabilities as

$$p_0(x) = \mathbb{P}(Y = 0|X = x), \quad p_1(x) = \mathbb{P}(Y = 1|X = x) = 1 - p_0(x), \tag{3.1}$$

and assign labels based on the estimated probabilities according to

$$h_{\mathcal{L}}(x) = \begin{cases} 0, & \text{if } p_1(x) < c \\ 1, & \text{if } p_1(x) \geq c \end{cases} \tag{3.2}$$

for a chosen probability cutoff $c \in [0, 1]$. For $c = 0.5$, this is equivalent to the assignment rule

$$h_{\mathcal{L}}(x) = \underset{k=0,1}{\operatorname{argmax}}\, p_k(x).$$

3.3 Evaluation

One of the main goals in classification (and in machine learning in general) is to construct a method that is generalizable and therefore can be reliably applied to new independent data, which was not used for training. Evaluating this quality of a classifier is of importance for the choice of the underlying model (e.g. parameter choices), but also for the performance assessment of the final classifier. In the following, we give an overview of typical measures used for evaluation purposes in machine learning. The error definitions as well as Section 3.3.3 are following notations from Hastie *et al.* (2009) (Chapter 7) and Louppe (2014), and Section 3.3.4 is based on Fawcett (2006).

3.3.1 Generalization error

We want to construct a classifier $h_{\mathcal{L}}$ in a way that it predicts reliably not only on the training set $\mathcal{L}$ but also on unseen data from $\mathcal{X} \setminus \mathcal{L}$. This can interpreted as minimizing the expected prediction error or **generalization error**

$$\mathrm{Err}(h_{\mathcal{L}}) = \mathbb{E}_{X,Y}[L(Y, h_{\mathcal{L}}(X))], \tag{3.3}$$

where the expected value is computed over all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and not only the measurements from $\mathcal{L}$. The function $L(Y, h_{\mathcal{L}}(X))$ measures the error between output Y and the predicted output $h_{\mathcal{L}}(X)$ and is called **loss function**.
A typical example in binary classification is the **0-1 loss function**

$$L(Y, h_{\mathcal{L}}(X)) = \mathbb{1}(Y \neq h_{\mathcal{L}}(X)) \tag{3.4}$$

which counts the number of misclassifications and penalizes each equally.
Unfortunately, the generalization error can rarely be computed directly since the distribution $\mathbb{P}(X, Y)$ is unknown. Instead, it has to be estimated from the data. The simplest estimate is the **training sample error** which is only based on $\mathcal{L}$ and can therefore be used when no additional data is available. It

is defined as the average loss over all samples in $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,

$$\widehat{\mathrm{Err}}_{\mathrm{train}} = \frac{1}{|\mathcal{L}|} \sum_{(x,y)\in\mathcal{L}} L(y, h_{\mathcal{L}}(x)).$$

As one can imagine, this is a rather poor estimate since a model trained on $\mathcal{L}$ is prone to make especially good predictions for these samples, but too optimistic estimates for unseen samples from $\mathcal{X} \setminus \mathcal{L}$. A better possibility would be to estimate the generalization error from an independent test set $\mathcal{L}' \subset (\mathcal{X} \times \mathcal{Y}) \setminus \mathcal{L}$ which has not been used for training and as such was not 'seen' by the classifier $h_{\mathcal{L}}$. Unfortunately, as mentioned earlier, the training set is often the only available labeled data. If the number of samples $n = |\mathcal{L}|$ is 'big enough', we can simply split $\mathcal{L}$ into non-overlapping training and test sets, $\mathcal{L} = \mathcal{L}_{\mathrm{train}} \cup \mathcal{L}_{\mathrm{test}}$. In this case, the so-called **test sample estimate** is defined as

$$\widehat{\mathrm{Err}}_{\mathrm{test}} = \frac{1}{|\mathcal{L}_{\mathrm{test}}|} \sum_{(x,y)\in\mathcal{L}_{\mathrm{test}}} L(y, h_{\mathcal{L}_{\mathrm{train}}}(x)), \tag{3.5}$$

and a reliable estimate of the generalization error of $\mathcal{L}$. If there are 'too few' samples in $\mathcal{L}$, the set used for training could be too small after putting aside the independent test set and both the trained classifier and the generalization error estimate might be of rather poor quality. An efficient sample re-use can be helpful in this situation. The two most commonly used methods are bootstrapping, which is for example used to construct a random forest classifier (see Section 3.4.3), and cross-validation, which we will explain in more detail in the following section

3.3.2 Cross-validation

Following the same idea as in the 'data rich' scenario, **cross-validation** uses a part of the data to train the model and another independent part to compute the test sample estimate defined in Equation (3.5). However, instead of looking at only two subsets of $\mathcal{L}$, the data is split into K parts of roughly equal size, $\mathcal{L} = \mathcal{L}_1 \cup \ldots \cup \mathcal{L}_K$. Then, for all $k = 1, \ldots, K$, we train a model $h_{\mathcal{L}\setminus\mathcal{L}_k}$ and compute the **average prediction error** over $\mathcal{L}_k$,

$$\overline{\mathrm{Err}}_{\mathrm{CV}_k} = \frac{1}{|\mathcal{L}_k|} \sum_{(x,y)\in\mathcal{L}_k} L(y, h_{\mathcal{L}\setminus\mathcal{L}_k}(x)). \tag{3.6}$$

In other words, we train on $K-1$ of the K subsets and use the remaining independent data part to compute the test sample estimate. The fact that every subset of the partition is once being left out for training comes with the advantage that we have predictions on all samples $(x, y) \in \mathcal{L}$. Also, the model $h_{\mathcal{L}\setminus\mathcal{L}_k}$ should be close to $h_{\mathcal{L}}$ as it was trained on a big proportion of $\mathcal{L}$.
The final estimate of the **generalization error for K-fold cross-validation** is computed by averaging over all K average prediction errors from Equation (3.6),

$$\widehat{\mathrm{Err}}_{\mathrm{CV}_k} = \frac{1}{K} \sum_{k=1}^{K} \overline{\mathrm{Err}}_{\mathrm{CV}_k}.$$

In most applications, either 5-fold or 10-fold cross-validation is used.

3.3.3 Model selection with parameter optimization

The classification model does usually not only depend on the training set $\mathcal{L}$ but also on free (tuning) parameters θ which describe the model complexity. Hence, before assessing the performance of a trained model $h_{\mathcal{L},\theta}$ on an independent test set, we first want to select the 'best' model configuration by finding the 'optimal' parameters θ. In principal, we would need another independent data set to measure the prediction errors. In a data-rich scenario, $\mathcal{L}$ can be split into a **training set**, a **validation set** for model selection and a **test set** for model assessment or evaluation. In the data-poor scenario, however, the training and model selection is usually done on the same set using, for example, cross-validation as described in Section 3.3.2, and only the model assessment is done on a beforehand separated independent test set.

3.3.4 Confusion matrix and associated performance metrics

In Section 3.3.1 we introduced the loss function as a necessary measure for classification performance where it represents the cost of a misclassification. Loss functions are chosen in a computational feasible way since they are part of the objective function used for model optimization, for example, when tuning model parameters θ as discussed in the previous Section 3.3.3. However, there are also other possibilities to assess an already optimized model. Many of the standard performance metrics are based on the classification outcomes summarized in the so-called **confusion matrix**, in which one dimension describes the actual class and the other the predicted class membership. For binary classification, this results in a 2×2 matrix as depicted in Table 3.1.

Table 3.1: Confusion matrix for binary classification problem.

	predicted positive	**predicted negative**
actual positive	true positive (TP)	false negative (FN)
actual negative	false positive (FP)	true negative (TN)

The first row describes the number of the measurements with a 'positive' outcome ($Y = 1$) which were correctly classified ($h_{\mathcal{L}}(Y) = 1$, true positives) or misclassified ($h_{\mathcal{L}}(Y) = 0$, false negatives), while the second row counts how many measurements with 'negative' outcome ($Y = 0$) are either misclassified ($h_{\mathcal{L}}(Y) = 1$, false positives) or correctly classified ($h_{\mathcal{L}}(Y) = 0$, true negatives). Probably the most prominent metric to measure classification performance from the results of a confusion matrix is **accuracy**, which is defined as the fraction of correctly labeled predictions (see Table 3.2). Accuracy values lie between 0 and 1, where a classifier with accuracy = 1 predicted all labels correctly. The accuracy measure has a one-to-one relationship to the 0-1 loss function (see Equation (3.4)) since the minimization of the 0-1 loss is equivalent to the maximization of accuracy.

When the test data is unbalanced, i.e., when the cardinality of one class is much bigger than of the other, accuracy is not the optimal metric to measure performance because it can give misleading results. If, for example, the class with output $Y = 0$ made up 90% of the data, then a model predicting exclusively this class label would have a high accuracy of 0.9, a 100% recognition rate for the bigger class, but a recognition rate of 0% for the smaller class corresponding to a true negative rate (TNR) of 1 and a true positive rate (TPR) of 0, respectively (see Table 3.2).

Table 3.2: Performance Metrics.

accuracy	$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$
recall, true positive rate	$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$
precision	$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$
false positive rate	$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$
specificity, true negative rate	$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$

The **receiver operating characteristics (ROC) curve** is another way of illustrating the performance of a binary classification task, but in contrast to the metrics before, it is based on the predicted conditional class probabilities defined in Equation (3.1) in Section 3.2. The ROC curve results from plotting the false positive rate (FPR) against the TPR for a gradually changing probability threshold and the corresponding class assignments according to Equation (3.2). The perfect ROC curve goes from $(0, 0)$ straight up to $(0, 1)$ and then straight to $(1, 1)$, and has an area under the curve (AUC) of 1. The expected AUC-ROC for a random class assignment is 0.5 independently of the trained model (see Figure 3.1). Just as accuracy, the results of the AUC-ROC can be misleading for unbalanced classification problems. The **precision-recall (PR) curve** is created by plotting precision against recall (equivalent to TPR) for class assignments based on varying probability thresholds, and is more informative in this situation (see Figure 3.1). The perfect curve starts in $(0, 1)$, goes straight to $(1, 1)$ and then down to $(f^*, 0)$, where f^* is the fraction of

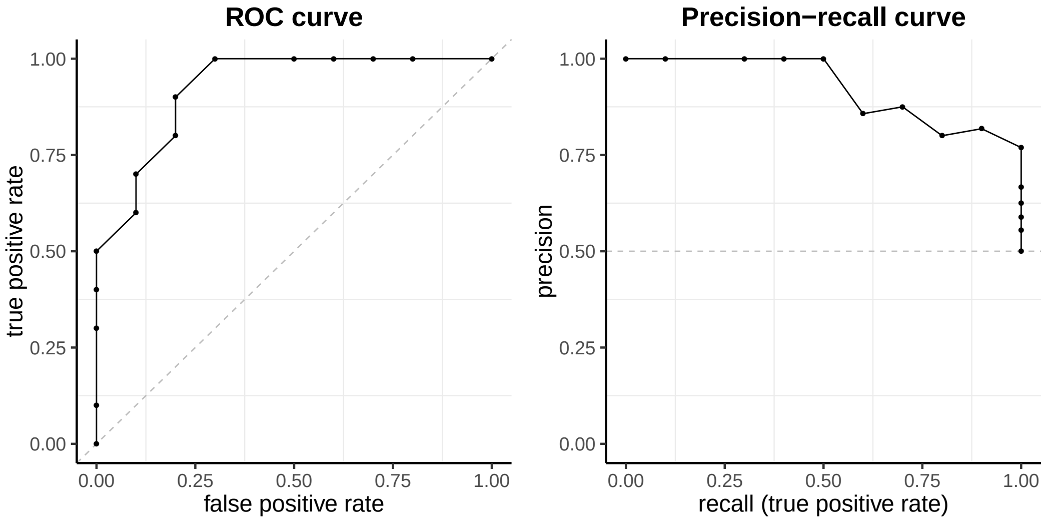


Figure 3.1: ROC curve and precision-recall curve. Performance results for predictions on 10 positive and 10 negative samples (balanced classification problem). Grey dotted lines indicate expected performance of a random classifier.

actual positives in the analyzed set of measurements. Thus, the area under the precision-recall curve (AUC-PR) is 1 at best. Furthermore, the AUC-PR corresponds to f^*, i.e., to the fraction of actual positives, for a random classifier.

3.4 Supervised learning

3.4.1 Introduction

In supervised learning, we are provided with a labeled training set $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ which consists of n realizations of the random vector $X = (X_1, \ldots, X_n)$ and the corresponding response variable Y. The goal is to learn a classifier $h_{\mathcal{L}} : \mathcal{X} \to \mathcal{Y}$ based on the training set $\mathcal{L}$ which can reliably predict the outcome $y \in \mathcal{Y}$ for a new data point x.

In this work, we are interested in **binary classification**, a subtype of supervised learning methods, where Y is a binary outcome variable, $Y \in \mathcal{Y} = \{0, 1\}$. The variety of available classification methods is huge, and the individual meth-

ods complex. Some of the most popular methods are linear discriminant analysis, logistic regression, support vector machines, the naive Bayes classifier, neural networks, decision trees and random forests (Hastie *et al.*, 2009).
In this work, we will focus on random forests, which we introduce according to Chapter 14 of Izenman (2008), unless stated otherwise. In order to understand the underlying principles of random forests, we also explain decision trees in more detail based on Chapter 2 of Breiman *et al.* (1984).

3.4.2 Decision trees

Decision trees are preferably used when the data has many (correlated) features and when the features influence the outcome in a complicated, nonlinear way. The main idea of tree-structured classifiers is to make complicated interactions of features more manageable by partitioning the measurement space $\mathcal{X}$ into smaller subsets. The construction of the tree starts by splitting $\mathcal{X}$ into two descendant subsets (see Figure 3.2 for an example). In a second step, both descendant subsets will be split into two even smaller subsets. This procedure is repeated until a set of (very) small regions is reached which are easy enough to handle to fit a simple model to them. The training set $\mathcal{L}$ is used to guide the construction of the decision tree. Each tree is divided in three main elements:

1. recursive partitioning/ determination of splits,
2. determination of when to stop splitting a subset, and
3. model fitting for each subset of the partition.

A decision tree comprises three types of core elements: tree nodes, branches and terminal nodes. Each **tree node** t represents a test that is performed on one of the p features of a measurement vector $x \in \mathcal{X}$, and the outcome of each test is represented by the **tree branches**. A **terminal node** represents the class label, in our case $l \in \{0, 1\}$, which is finally assigned by the tree to each cell of the final partition. The class prediction of a measurement vector x is determined by its path through the tree and corresponds to the class label of the terminal node in which it lands.

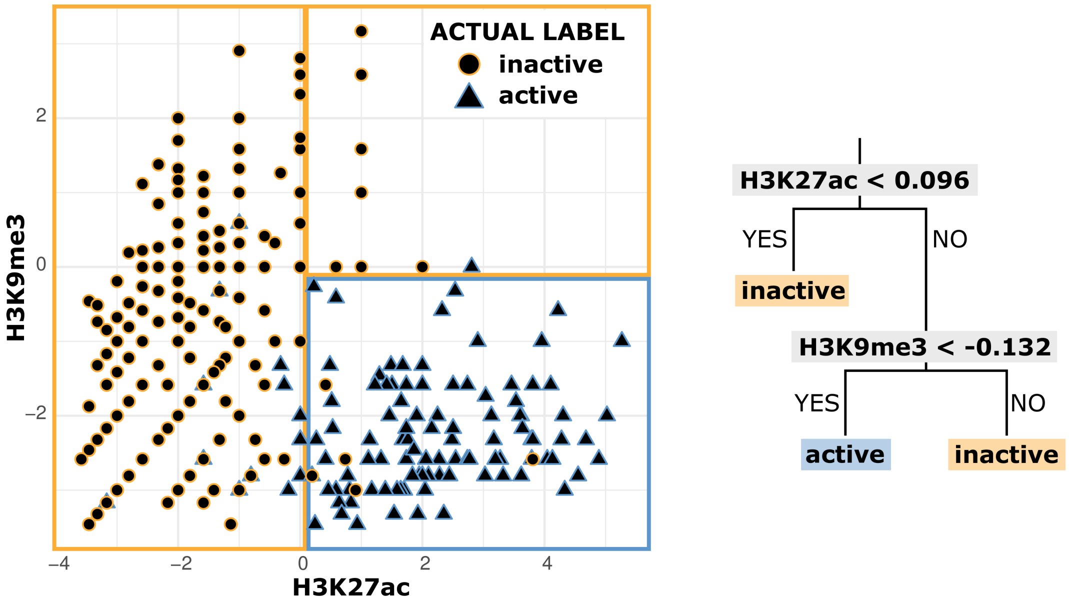


Figure 3.2: Example of a binary decision tree distinguishing 'active' from 'inactive' genomic regions. We trained a decision tree based on input-normalized histone modification ChIP-seq data to distinguish 'active' from 'inactive' genomic regions. The tree has two splits based on H3K27ac and on H3K9me3, and three terminal nodes representing the terminal partitions with labels. We plot normalized counts of H3K9me3 and H3K27ac for the 'active' (blue triangle) and 'inactive' (orange point) regions in the training set and include the partitions as rectangular shapes where the colour corresponds to the predicted label. The majority of regions would have been classified correctly based on the decision tree (i.e., blue samples in blue rectangle, orange samples in orange rectangle).

It turns out that the model fitting, i.e., the assignment of labels to the terminal nodes, is a simple problem, whereas the determination of splits and the decision to terminate the splitting process are much more challenging.

The fundamental idea of the recursive partitioning is that after each split the data becomes more homogeneous (in terms of class membership) in each of the two descendant regions than it was in the parent region. We define the node proportion $p(l, t)$ as the proportion of training measurements $x_i \in \mathcal{L}$ which are located in the the subset $R_t \in \mathcal{X}$ associated to node t, and additionally belong to class $l \in \{0, 1\}$:

$$p(l,t) = \frac{1}{|\{x_i \in R_t\}|} \sum_{i:\, x_i \in R_t} \mathbb{1}(y_i = l).$$

Thus, for every node t it has to hold that $p(0,t)+p(1,t) = 1$. The node proportion is subsequently used to compute a measure of homogeneity or ‘pureness’ for each node t, called impurity measure. The **node impurity** $i(t)$ can be computed in different ways, for example using entropy, the misclassification error, or the Gini index. These are all non-negative functions that give higher values when the class labels of training samples in R_t are very mixed and smaller values in case of a homogeneous class membership. The most commonly used rule to create the partitioning is the **Gini index**, which is defined as

$$i(t) = 1 - \sum_{l=0}^{1} p(l,t)^2 = 1 - p(0,t)^2 - p(1,t)^2$$

in the 2-class problem. The Gini index of a pure subset R_t is zero because one of the squared node proportions would be one and the other zero, and it reaches its maximum value when both of the proportions are equal, $p(0,t) = p(1,t) = 0.5$. The goodness of a split is now determined by the decrease in impurity. Let ρ be the binary split that divides a parent region R_t into two descendent regions R_{t_1} and R_{t_2}, and let $p(R_{t_1}) = \frac{|\{x_i \in R_{t_1}\}|}{|\{x_i \in R_t\}|}$ and $p(R_{t_2}) = \frac{|\{x_i \in R_{t_2}\}|}{|\{x_i \in R_t\}|}$ be the corresponding proportions of training data in R_t moved to subset R_{t_1} or R_{t_2}, respectively. Then, the goodness of split ρ is computed as

$$\Delta i(\rho, t) = i(t) - p(R_{t_1}) i(t_1) - p(R_{t_2}) i(t_2).$$

The maximum decrease of impurity in the scenario of a perfect split into two pure descendent regions R_{t_1} and R_{t_2} is equal to the Gini index of R_t, since $i(t_1) = i(t_2) = 0$ and hence $\Delta i(\rho, t) = i(t)$.

At every node t a certain set $\mathcal{R}$ of binary splits ρ is defined for every feature X_j, $j = 1, \ldots, p$. How the set $\mathcal{R}$ is chosen depends on the nature of the feature and the values represented in the training set. In our case, the features are continuous and the number of observations x_i in the training set gives the upper boarder of possible binary splits, $n-1$, for each feature. The split ρ with the maximum decrease in impurity is chosen, but it is only realized in case it fulfills a heuristic criterion which can be based on a threshold for $\Delta i(\rho, t)$ or on a maximum number of observations that have to fall into each descendent subset R_{t_1} and R_{t_2}. If at a node t non of the splits fulfill the criterion, t becomes a terminal node. The model fitting in the case of classification, like already mentioned, is rather simple. A terminal node t will be assigned the class label $l_t \in \{0,1\}$ which is most abundant in the corresponding subset of the partition R_t:

$$l_t = \underset{l \in \{0,1\}}{\operatorname{argmax}} \sum_{i:\, x_i \in R_t} \mathbb{1}(y_i = l)$$

where y_i, $i = 1, \ldots, n$, are the output measurements or labels from the training set $\mathcal{L}$.

The predicted class membership for an input measurement $x \in \mathcal{X}$ is equal to $h_{\mathcal{L}}(x) = l_t$, where l_t is the label of the terminal node that includes x.

3.4.3 Random forest

A random forest classifier is a learning method consisting of a collection of decision trees. Each individual tree is based on a different randomly sampled subset of the training observations, and each split within a tree is determined from a randomly sampled subset of features. The final output or prediction for an observation vector $x \in \mathcal{X}$ is a probability value resulting from the fraction of trees voting in favor of the class of interest (see also Figure 3.3).

Let the random forest classifier consist of an ensemble of $m = 1, \ldots, M$ randomized tree-structured classifiers $h_{\mathcal{L}}(\theta_m) : \mathcal{X} \rightarrow \{0,1\}$ based on independent identically distributed random vectors $\{\theta_m\}_{m=1}^{M}$. More precisely, for each tree T_m a random vector θ_m is generated which is independent of the $m-1$ previously generated vectors $\{\theta_1, \ldots, \theta_{m-1}\}$. According to θ_m, a bootstrap sample $\mathcal{L}_m$ of size n is drawn with replacement from the training set $\mathcal{L}$ to construct the decision tree T_m. Each split in the tree is based on only a small subset of the available p features. The number of chosen features p' can vary, but is per default set to $p' = \sqrt{p}$. Finally, the trees are growing without pruning, i.e.,

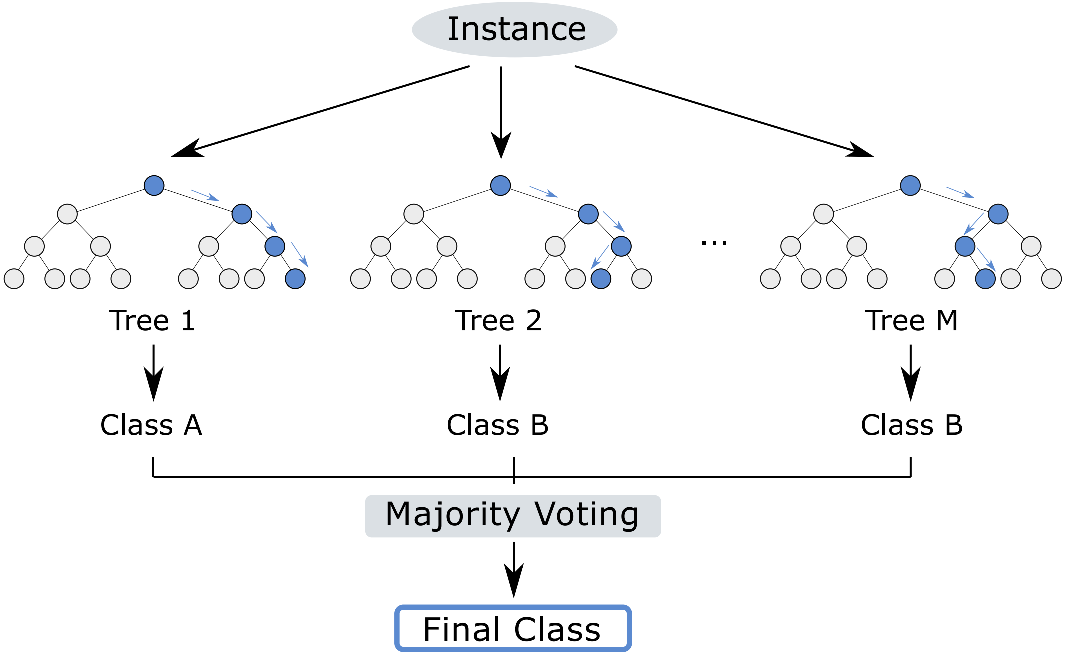


Figure 3.3: Classification with random forest. A new data instance is labeled by each of the M decision trees according to its path through the tree and its terminal node. The final class membership of the new data instance is based on the majority voting of the individual tree labels.

they are grown to their maximum size. The individual steps of constructing a random forest classifier are summarized in Algorithm 1.

The final output of a random forest for an input measurement $x \in \mathcal{X}$ is the predicted conditional class probability which is computed from the voting of the individual decision trees according to

$$p_l(x) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(h_{\mathcal{L}}(x, \theta_m) = l), \quad l \in \{0, 1\}.$$

The class membership of x can then be assigned according to a cutoff c as explained in Equation (3.2) in Section 3.2.

```
Input: learning set $\mathcal{L}$, number of trees $M$, number of features to be
  chosen at each split $p'$
Output: random forest as a collection of $M$ tree-structured classifiers
  $\{h(x, \theta_1), \ldots, h(x, \theta_M)\}$
  for $m = 1, \ldots, M$ do
  • generate random vector $\theta_m$ independent of $\{\theta_1, \ldots, \theta_{m-1}\}$
  • draw bootstrap sample $\mathcal{L}_m$ from $\mathcal{L}$ according to $\theta_m$
  • grow a tree $T_m$ based on $\mathcal{L}_m$ using random feature selection
      (at each node randomly select $p'$ features from $X = (X_1, \ldots, X_p)$
      according to $\theta_m$ and find best split based on Gini index)
  • define classifier $h(x, \theta_m)$ having a single vote for the class of $x \in \mathcal{X}$
  end for
```

Algorithm 1: Random forest

Advantages of 'random components'

Random forests are characterized by the combination of two 'randomizing components', bootstrap aggregation and random feature selection, which together create a collection of decision trees. Each single tree by itself would not be an optimal classifier. Growing until full depth without pruning, the trees are expected to have a low bias and a high variance, which means that they are prone to overfitting to their training data. However, each tree is based on different training data obtained by bootstrap aggregation (drawing by replacement) and additionally, each split in the tree is based on a randomly selected subset of features. This leads to a decrease in dependence or correlation between the individual trees, and therefore in a lower generalization error of the forest (see Equation (3.3) in Section 3.3.1 for more details on generalization error). In fact, it was shown by Breiman (2001) that the generalization error of a random forest converges almost surely to an upper bound with increasing number of trees M ($M \to \infty$) in dependence of only two parameters, the strength or accuracy of each individual tree and the correlation between the trees. The existence of such an upper bound shows that random forests are not prone to overfitting especially for a large number of trees. Another practical advantage of bootstrap aggregating is that it can be used to make ongoing out-of-bag estimates of the error, strength and correlation of the trees in the forest without the need of a set-aside labeled test set.

Also, in case of high correlation between features, the random feature selection can be useful to acknowledge the importance of each feature individually. It still holds that once a feature is used in the tree, a highly correlated one has a decrease in importance, but since they will not always be chosen together in a tree, both features should in theory play a role in the tree construction. Finally, the random feature selection step results in a decrease in computational time compared to an evaluation of all available features.

3.5 Unsupervised learning

3.5.1 Introduction

In unsupervised learning, we are provided with learning or training data which only consists of unlabeled input measurements $\mathcal{U} = \{x_1, \ldots, x_n\}$ as realizations of the random vector $X = (X_1, \ldots, X_P) \in \mathcal{X}$. The goal is to get a better understanding of the data by exploring its underlying (hidden) structure. While for low-dimensional problems like $p = 1, 2, 3$ there are methods to directly compute certain properties of the data, the task gets much harder in higher dimensions. In fact, the number of features p and as such the dimensionality of the data can be quite high in many unsupervised settings, often much higher than for supervised learning. This leads to an estimation problem since it becomes more likely that the available learning data $\mathcal{U}$ is sparse in the inflated feature space and as a result not every feature is sufficiently represented.

Two well known strategies to tackle this problem are clustering and dimensionality reduction. **Dimensionality reduction** aims at reducing the data space by inferring a small set of so-called latent variables from the observed features. Based on these latent variables, the data can be projected (and handled) in a smaller space without loosing too much of the original information. Principal component analysis is one of the most prominent examples for dimensionality reduction. **Clustering** methods aim to find informative patterns or clusters of similar data with potentially simpler underlying structures. However, the applied similarity measures or cost functions can vary between different algorithms which may lead to different clustering results. Also, the number of desired clusters often needs to be specified beforehand (Duda *et al.*, 2001). **Hidden Markov models** (HMMs) are also common methods to infer latent variables. They do not constitute a classical example of unsupervised methods since they can be used for classification as well given the necessary labeled data. However, HMMs can be learned in an unsupervised way using the **expectation maximization** (EM) algorithm to infer the latent or hidden variables.

In this work, we focus on HMMs as an example of unsupervised methods since they form the basis of several enhancer prediction methods and will be discussed in this context in Chapter 4. Before we explain HMMs in more detail, we will introduce the concept of **Markov Chains** in the following chapter. The presented material is mainly based on the Chapters 7 and 10 of Koski (2001), on Chapter 3 of Durbin *et al.* (1998) and Chpater 8 of Hastie *et al.* (2009).

3.5.2 Markov chains

A sequence of random variables $\{S_t\}_{t\in\mathbb{N}_0}$ taking values in a finite set $\mathcal{S} = \{\varsigma_1,\dots,\varsigma_J\}$ is called **Markov chain** (of order 1), if for all $t \geq 1$ and for all $s_0, s_1, \dots, s_t \in \mathcal{S}$ there holds

$$\mathbb{P}(S_t = s_t | S_0 = s_0, \dots, S_{t-1} = s_{t-1}) = \mathbb{P}(S_t = s_t | S_{t-1} = s_{t-1}).$$

This condition is also called **Markov property** and says that the future (t) and the past states ($t-2$, $t-3$, ...) of the chain are independent if conditioned on the present state ($t-1$). In other words, the chain ‘has no memory’.

The conditional probabilities

$$p_{i,j}^{(t)} = \mathbb{P}(S_t = \varsigma_j | S_{t-1} = \varsigma_i), \quad t \geq 1, \quad i,j = 1,\dots,J$$

are called one step **transition probabilities** and $P^{(t)} = (p_{i,j}^{(t)})_{i,j=1,\dots,J}$ is the $J \times J$ **transition (probability) matrix**. We consider time-homogeneous Markov chains, for which the transition probabilities are independent of the time, $p_{i,j}^{(t)} = p_{i,j}$, and hence $P^{(t)} = P$ for all $t \geq 1$. Since the i-th row of the transition matrix P describes the conditional probability distribution of S_t given the knowledge that the chain was in state ς_i in the immediate past, $S_{t-1} = \varsigma_i$, it holds true that

$$p_{i,j} \geq 0 \quad \text{for all } i, j = 1, \ldots, J \quad \text{and} \quad \sum_{j=1}^{J} p_{i,j} = 1.$$

The **initial distribution** is defined by $\pi(0) = (\pi_{\varsigma_0}(0), \ldots, \pi_{\varsigma_J}(0))$ with $\pi_{\varsigma_j}(0) = \mathbb{P}(S_0 = \varsigma_j)$ describing the probability of starting in state ς_j at time $t = 0$ for $j = 1, \ldots, J$. A Markov chain is uniquely defined by its initial distribution $\pi(0)$ and the transition probabilities in P.

3.5.3 Hidden Markov models

Markov chains are defined in a way that the state of the random sequence at time t is directly observable, since state sequence and output live in the same space $\mathcal{S}$ and have a one-to-one correspondence. For hidden Markov models (HMMs), however, the underlying state of the model is hidden and as such cannot be directly inferred from the output. Instead, an HMM can be characterized by

1. defining a hidden Markov chain,
2. defining an observable random process, and
3. the assumption of conditional independence between the observations and the hidden state sequence.

The **hidden Markov chain** is a sequence of random variables $\{S_t\}_{t \in \mathbb{N}_0}$ which takes values in $\mathcal{S} = \{\varsigma_1, \ldots, \varsigma_J\}$ and follows a (one-step) time-homogeneous Markov chain characterized by an initial distribution $\pi(0)$ and a transition probability matrix P with properties as described in Section 3.5.2.

Since the output of the model is no longer a direct result of the (hidden) sequence states modeled by $\{S_t\}_{t \in \mathbb{N}_0}$, an additional sequence of random variables is required for modeling the HMM. The **observable random process** $\{O_t\}_{t \in \mathbb{N}_0}$, lives in the finite state space $\mathcal{O} = \{\omega_1, \ldots, \omega_K\}$, where K does not have to be equal to J. The two processes $\{S_t\}_{t\in\mathbb{N}_0}$ and $\{O_t\}_{t\in\mathbb{N}_0}$ are connected through the so-called **emission probabilities**

$$e_{\varsigma_j}(\omega_k) = \mathbb{P}(O_t = \omega_k | S_t = \varsigma_j), \quad j = 1, \dots, J, \quad k = 1, \dots, K,$$

for a fixed $t \geq 1$. In words, $e_{\varsigma_j}(\omega_k)$ describes the probability that output $\omega_k \in \mathcal{O}$ is observed when the model is in state $\varsigma_j \in \mathcal{S}$. For the emission probabilities there holds

$$e_{\varsigma_j}(\omega_k) \geq 0 \quad \text{for all } j = 1, \dots, J;\ k = 1, \dots, K \quad \text{and} \quad \sum_{k=1}^{K} e_{\varsigma_j}(\omega_k) = 1.$$

They can be summarized in the matrix $E = (e_{\varsigma_j}(\omega_k))_{j=1,k=1}^{J,K}$. Furthermore, we assume **conditional independence** of the emitted output sequence $o = (o_0, \dots, o_t)$ given the corresponding vector of hidden states $s = (s_0, \dots, s_t)$:

$$\mathbb{P}(O = o | S = s, E) = \prod_{i=0}^{t} e_{s_i}(o_i),$$

with $O = (O_0, \dots, O_t)$ and $S = (S_0, \dots, S_t)$. As a consequence of the introduced assumptions, we can write the joint probability of an emitted output sequence o and the hidden states s as

$$\mathbb{P}(O = o, S = s) = \pi_{s_0}(0) e_{s_0}(o_0) \prod_{i=1}^{t} p_{s_{i-1},s_i} e_{s_i}(o_i).$$

Computing the joint probability can be challenging, since we often do not know all variables needed. In the unsupervised applications of HMMs it is actually the main goal to infer the hidden states of the model from the given data.
Given the case that we observe a sequence $o = (o_0, \dots, o_t)$ and we know the initial, emission and transition probabilities/parameters, we have two possibilities to infer or decode the underlying hidden states $(s_0, \dots, s_t)$, either Viterbi coding or posterior coding (forward-backward algorithm). Since this is not the focus of this work, we will not go into further details but refer the interested reader to Durbin *et al.* (1998) for more information.

In the unsupervised scenario, which we are interested in, the only given parameter is the number $|\mathcal{S}| = J$ of hidden states, which is probably the most difficult HMM scenario. All other parameters have to be inferred based on given observations, $o = (o_0, \ldots, o_t)$, which can be done by maximizing the likelihood function of the observed data,

$$\mathbb{P}(O = o; E, P, \pi) = \sum_{s \in \mathcal{S}^{t+1}} \mathbb{P}(O = o, S = s; E, P, \pi).$$

Since there is no closed-form solution for this problem, this has to be done by estimation. The **Baum-Welch algorithm** is used as a standard in the context of unsupervised HMMs. It integrates the already mentioned forward-backward algorithm into the well-known **expectation maximization** (EM) algorithm. The EM approach, an iterative algorithm to perform maximum likelihood estimation, is widely used in the context of HMMs since it is applicable when models depend on latent or hidden parameters. Let θ be the entire set of parameter values $\{E, P, O\}$ used in this model. The EM algorithm starts from an initial arbitrary choice of parameters, $\theta^{(0)}$, and improves them by repeatedly applying an expectation (E) and a maximization (M) step. More precisely, an improved parameter $\theta^{(t+1)}$ is computed in every iteration step t by maximizing the expected log-likelihood function of the data evaluated with the current estimate $\theta^{(t)}$:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \sum_{s \in \mathcal{S}} \mathbb{P}(S = s | O = o, \theta^{(t)}) \log \mathbb{P}(O = o, S = s, \theta).$$

The forward-backward algorithm is needed in the expectation step to infer the underlying hidden states from the data and the current estimates $\theta^{(t)}$. The iteration stops when some convergence criterion is reached.

4 Enhancer Prediction Method Based on Histone Modification Data and a Combination of Two Random Forest Classifiers

In this chapter, we describe the theoretical background of a new supervised method to predict the location of enhancers genome-wide solely based on histone modification ChIP-seq data. On the example of a mouse embryonic stem cell (mESC) data set, we present the parameter optimization strategy used to create our final enhancer classifier. In this context, we also assess the quality of the mESC samples and the enhancer set, which we chose for training. Finally, we present in detail how we perform motif analysis on predicted enhancer regions, and we introduce the theoretical background of two other methods commonly used for enhancer prediction.
Our enhancer prediction method presented in this chapter is an extended and adapted part of a paper which was co-first authored with Verena Heinrich (Ramisch *et al.*, 2018). The content of this chapter is based on my part of the collaboration.

4.1 Framework of our prediction method

Genome-wide enhancer prediction describes the task to decide for each base pair position in the genome if it encodes for an active enhancer or not based on one or more chosen measurable features. As one can imagine, this decision is of varying difficulty for different positions depending on the real underlying genomic element (e.g. promoters, genes, intergenic regions) and on how well the chosen feature set is suited to contrast enhancers from each type of genomic element.

The way we tackle this problem is to segment the genome into equally sized bins of 100 bp each, predict the enhancer probability

$$\mathbb{P}(\text{bin}_x = \text{active enhancer})$$

for each bin_x using a supervised approach and assign an 'enhancer' or 'non-enhancer' label to the corresponding bin_x based on the predicted probability. The supervised classification model has to be trained on a set of already labeled input data, the so-called training set, on which the class-specific feature patterns are learned (see Section 3.4 for more details on supervised classification).

Our prediction model is a combination of two random forest classifiers with histone modification-based feature sets, where one classifier learns the difference between 'active' and 'inactive' genomic regions, and the other learns to distinguish active enhancers from active promoters. Each random forest consists of several decision trees which determine the final classification outcome via majority voting (see also Section 3.4.3). The training enhancers are chosen according to their level of bidirectional transcription and accessibility.

The successfully trained model can predict enhancers genome-wide solely based on ChIP-seq data from six core HMs and a control experiment. The integrated implementation of the two random forest classifiers results in two genome-wide probability tracks (see Figure 4.1). These are multiplied for each genomic bin to generate the final enhancer probability values, based on which a list of annotated enhancers is created. Moreover, closely positioned enhancers are summarized to enhancer clusters.

In the following sections, we explain the individual steps of building our enhancer prediction model. First, we illustrate the feature choice, followed by the idea to combine two classification tasks to handle the heterogeneous 'non-

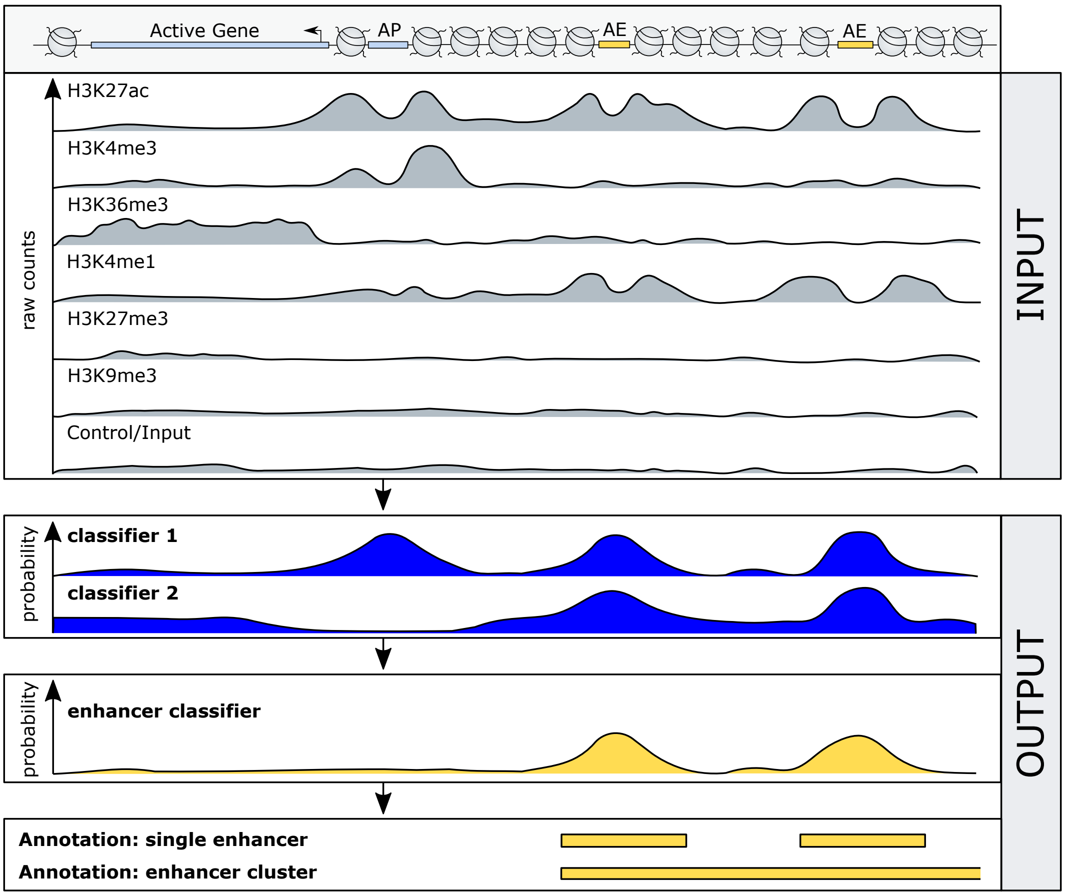


Figure 4.1: Framework of our enhancer prediction method. The two pre-trained classifiers (classifier 1 and 2) are applied to ChIP-seq data and result in two genome-wide probability tracks (dark blue). Classifier 1 predicts the probability of an 'active' region, and peaks at the active promoter ('AP', small rectangle, light blue) and both active enhancers ('AE', small rectangle, yellow). Classifier 2 distinguishes active enhancers from active promoter, and reaches high probabilities at the active enhancers and a low probability at the active promoter. Multiplying classifier 1 and 2 results in genome-wide enhancer probabilities (yellow), which are used to create enhancer annotations (small rectangles, yellow). These are clustered into regions of high enhancer density (broad rectangle, yellow). ChIP-seq count distributions are illustrative.

enhancer' set in the genome. Then, we discuss the training set composition, the necessity to choose training and feature set in an independent manner, as well as how we define enhancers and promoters for training in detail. Finally, we demonstrate how to annotate individual enhancer peaks from the predicted genome-wide enhancer probabilities, and how to combine or cluster these peaks into regions of high enhancer density.

4.1.1 Feature choice

The nucleosomes flanking an active enhancer show specific histone modification (HM) patterns which change dynamically according to the activity status of the enhancer (see Section 2.4.2 for more details). Therefore, HM data is a suitable and prominent feature candidate for (cell type-specific) enhancer prediction. The feature set for our enhancer prediction is derived from six core histone modifications: H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3. These marks are available for many different cell types and tissues as they make up the minimum set of core HMs defined by the International Human Epigenome (IHEC) consortium (Stunnenberg *et al.*, 2016), and are also included in the NIH Roadmap Epigenomics Mapping Consortium (Bernstein *et al.*, 2010). The challenge is to collapse the information of these HM marks at each genomic location into one informative measure of enhancer activity, which is in our case an enhancer probability track.

We compute the raw ChIP-seq read counts in each non-overlapping 100 bp bin for the respective HM feature sets and the control or input sample using the R package *bamProfile* (Mammana and Helmuth, 2016). Then, we perform an input normalization by computing the $\log_2$ ratio between the counts of each HM and the input (both with pseudocount of 1) per bin.

To account for the physical structure of an enhancer (an accessible region flanked by nucleosomes), the prediction for each genomic bin depends also on the HM read counts in multiple adjacent bins.

As shown in Figure 4.2, we expect a very low HM signal at the accessible nucleosome-free center of an active enhancer, the 100 bp bin$_x$. For enhancer associated marks like H3K27ac and H3K4me1 we expect to see strongly increas-

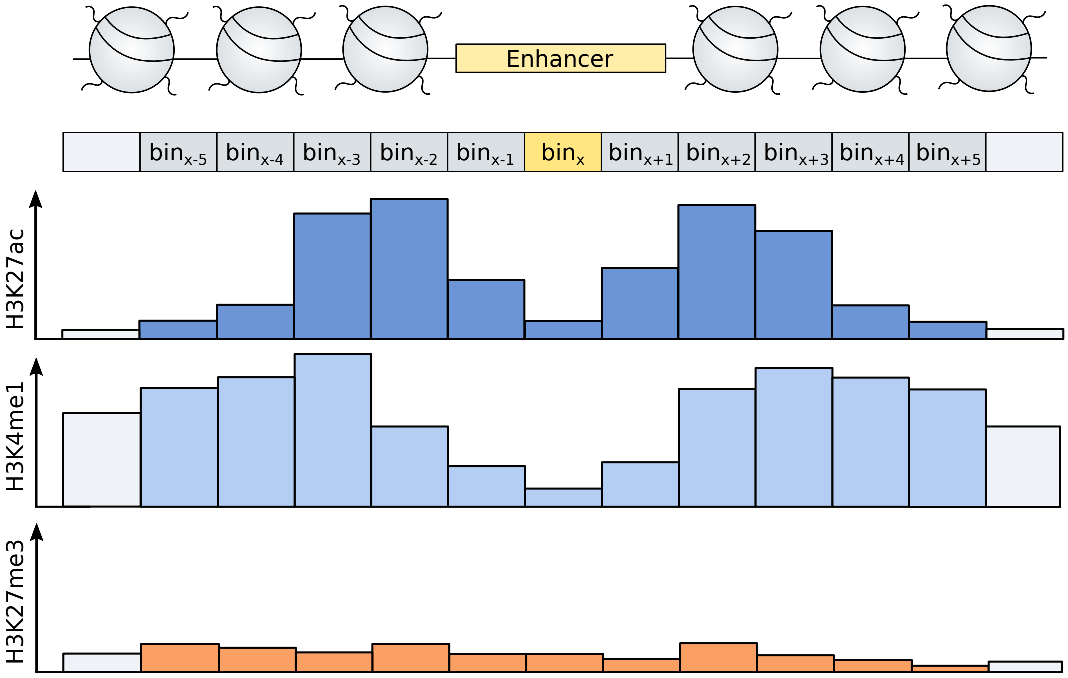


Figure 4.2: Illustrative HM profile at an active enhancer region. Count signals (on the y-axis) of H3K27ac, H3K4me1 and H3K27me3 at the center of an active enhancer (bin$_x$, in yellow) and five adjacent bins up- and downstream (in grey). Low HM signals at the accessible region and peaks for enhancer-associated marks H3K27ac and H3K4me1 at the flanking nucleosomes. Overall low signal for repressive mark H3K27me3.

ing signals in a number of neighbouring 100 bp bins to the left (bin$_{x-1}$, bin$_{x-2}$, . . .) and to the right (bin$_{x+1}$, bin$_{x+2}$, . . .) reflecting the level of modifications at the flanking nucleosomes, while for repressive marks like H3K27me3 we do not expect to see any signal in the vicinity of the active enhancer. Taking into account N neighboring bins to both sides of bin$_x$, this results in $2N + 1$ features (normalized count values) per HM. The number N of neighboring bins considered in the model is a parameter that is optimized as described in Section 4.3.

4.1.2 Combination of two random forest classifiers

According to the genome-wide distribution of several HM marks only a small fraction of the genome (about 10%) has enhancer potential (Kellis *et al.*, 2014). This results in an unbalanced classification problem where the class of interest is underrepresented. Furthermore, the remaining part of the genome is very diverse. It consists of active and inactive promoters, corresponding genes including exons and introns, and other types of genomic elements. Based on prior knowledge about properties of different genomic elements, e.g., through the occurrences of certain HMs throughout the genome, it can be expected that the distinction of active enhancers and 'non-functional background' regions is rather easy, while active enhancers and active promoters are much harder to distinguish (see Section 2.4). Hence, it is desirable to learn individual strategies how to distinguish enhancers from different types of genomic regions instead of considering one very heterogeneous 'non-enhancer' class.
Since the probability of observing an active enhancer can also be written as

$$\begin{aligned} P(\text{bin}_x =&\text{active enhancer}) = \\ &\underbrace{\mathbb{P}(\text{bin}_x = \text{active})}_{\text{classifier 1}} \cdot \underbrace{\mathbb{P}(\text{bin}_x = \text{active enhancer} \,|\, \text{bin}_x = \text{active})}_{\text{classifier 2}}, \end{aligned}$$

we split the task of enhancer prediction into two individual classification problems. The first classifier, **classifier 1**, learns to distinguish active from inactive genomic regions, while the focus of **classifier 2** lies in the distinction of active enhancers and active promoters. In case a region does not show any active features, classifier 1 should predict a very low probability and as a result, the final enhancer probability will be low as well independent of the outcome of classifier 2. If classifier 1 assigns a high probability, the enhancer decision is transferred to classifier 2. The combination, i.e., the product, of both classifiers results in the final enhancer prediction and is summarized in Figure 4.1.
Classifier 1, which has the objective to distinguish active regions from the rest of the genome, is based on a subset of the six modifications: the active mark H3K27ac and the repressive marks H3K27me3 and H3K9me3. We consider all six HMs as well as the ratio between H3K4me1 and H3K4me3 to learn

the difference between active promoters and enhancers with **classifier 2**. The H3K4me1/H3K4me3 ratio is computed as the $\log_2$ ratio between the input normalized values of H3K4me1 and H3K4me3 per bin (with pseudocount of 1), after shifting their distributions to the positive numeric values (> 0). It is expected to be higher at enhancers and was found to be an important feature (Calo and Wysocka, 2013).

For both classifiers, the features are correlated not only due to the interplay of individual HMs (Lasserre *et al.*, 2013), but especially due to their (neighboring) genomic location, e.g., the level of H3K27ac at bin_x is expected to be correlated to the level of H3K27ac at bin_{x-1} or bin_{x+1}.

We use **random forest** algorithms for both classification tasks since random forests are known to show a robust performance under a correlated feature assumption. In addition, they can provide us with information about the importance of the individual features for the classification task. A random forest is an ensemble classification method consisting of a number M of decision trees. For a genomic bin_x, each individual tree votes for a class membership and the enhancer probability is the ratio of all positive voting outcomes. More details on decision trees and random forests can be found in Sections 3.4.2 and 3.4.3, respectively. We optimize the number of decision trees M for both classifiers independently as described in Section 4.3.

4.1.3 Training and test set composition

The **training set** of a classification task comprises a **positive set** and a **negative set** of labeled data from which the classifier can adequately learn differentiating feature patterns. The individual compositions of the positive and negative set, i.e., the fraction and type of positive and negative examples, are important to give a complete summary of the class of interest regarding the chosen feature set and to create a representative background class.

For our model, we train two random forest classifiers and hence also require two different training sets. **Classifier 1** learns to distinguish between active genomic regions (active enhancers and active promoters) constituting the positive set and inactive genomic regions (intergenic, intragenic and inactive promoters) as the negative set. The proportion of the individual regions in the training set takes into account the composition of the genome as described in Kellis *et al.* (2014). We chose 10% active enhancers, 2% active promoters, 2% inactive promoters, 6% intragenic and 80% intergenic regions, summing up to 1000 regions in total. **Classifier 2** learns the difference between the two types of active regions used in classifier 1, i.e., the active enhancers constitute the positive set and active promoters the negative one. It is trained on 120 regions of which $83.\bar{3}\%$ are active enhancers and $16.\bar{6}\%$ are active promoters, keeping the same promoter-enhancer ratio and total numbers as used in classifier 1.

We decided to reflect the genome-wide imbalance between enhancers and non-enhancer regions to not introduce a deliberate sample selection bias on the training set. That means, we keep the (label) distribution of the training set as similar to the real genome-wide distribution as possible to avoid the need of making adjustments on the prediction results. It is know that without making adjustments on new data predictions, which were based on shifted prior probabilities in the training set, likely results in a loss of performance in comparison to predictions that took the real underlying prior class probabilities into account during training (Amos, 2008; Latinne *et al.*, 2001).

The performance of the overall prediction, which results from a combination of the two classifiers, is computed on an independent **test set** of 1000 examples. Since we are interested in how well active enhancers can be predicted, the positive set contains only enhancers while the negative set comprises the remaining genomic regions. The proportion of individual elements is kept as for classifier 1, i.e., the positive class makes up 10% of the test set.

4.1.4 Independence of training and feature set

If we would be provided with a large set of *in vivo* tested enhancers, we could train a classifier based on HM data (or any other feature of interest) without being concerned about potential introduced biases, since the training enhancers would be chosen independently of their individual HM level.

However, in most cases there are not enough *in vivo* tested enhancers available in the cell type of interest, and observed enhancer probabilities have to be exploited to create a reliable positive set. As a consequence, the next challenge is to construct a training set which is ideally completely independent of the classification features to avoid circular reasoning.

Example of strongly correlated feature and training criterion

Let us assume, we take promoter distal H3K27ac peaks as a proxy for enhancers in our training set, which is common in certain applications (e.g. Sun *et al.* (2016) and Novo *et al.* (2018)). For the construction of this particular training set, we would have to use two kind of a priori observed knowledge: a cutoff on the peak height of H3K27ac, c_{H3K27ac}, which distinguishes active enhancers ($Y = 1$) from background ($Y = 0$), and a distance cutoff, c_{dist}, distinguishing enhancers from promoters ($Y = 0$). Assuming we would use H3K27ac peak height and promoter distance also as features, is there additional information hidden in the data that the classifier could learn, which we did not already know during the training set design? The answer is no, since the classifier will achieve the best performance results following our own design rules:

$$Y = \begin{cases} 1, & \text{if H3K7ac} \geq c_{\text{H3K27ac}} \text{ and promoter distance } \geq c_{\text{dist}}, \\ 0, & \text{if H3K7ac} \geq c_{\text{H3K27ac}} \text{ and promoter distance } < c_{\text{dist}}, \\ 0, & \text{if H3K7ac} < c_{\text{H3K27ac}}. \end{cases}$$

If we would replace H3K27ac with a very correlated feature, like H3K9ac, the classification rules would change, but only as a function of the correlation between the two HMs. Hence, by using correlated training set criteria and features, we bias our predictions towards self-set rules and also challenge the need and gain of training a classifier in the first place.

Suitable training criteria for HM-based feature set

Considering that we want to use HMs as features, we want to choose enhancer properties that are most suitable and at the same time available for the construction of a reliable training set. In Section 2.4, we discussed many observed properties in detail: the binding of cell type-specific TFs and other co-factors, functional and sequence conservation, accessibility, bidirectional transcription, DNA methylation, and also enhancer-promoter communications within TADs. There are several classifiers which rely on distal p300 binding sites as a proxy for enhancers and use HMs in their feature set, for example Rajagopal *et al.* (2013) and He *et al.* (2017). Since p300 is a histone acetyltransferase (HAT) and as such is directly correlated to the presence and level of H3K27ac, we excluded p300 binding sites as valid criterion for our training set. In addition, it was found that there is a subset of enhancers in human ESCs which are preloaded by p300 while they are in a poised state (Rada-Iglesias *et al.*, 2011). These regions would enter the training set as false positives.

Some classifiers, especially earlier ones, rely solely on conservation scores (Siepel *et al.*, 2005; Prabhakar *et al.*, 2006). We did not include these features into our analysis since many enhancers show very low conservation levels even between closely related species (Schmidt *et al.*, 2010; Blow *et al.*, 2010; May *et al.*, 2012). Furthermore, we want to predict enhancers in different cell types and tissues, where different TFs may be important for enhancer activity. Hence, introducing conservation into the training set might bias the analysis towards a specific set of enhancers.

4.1.5 Definition of training enhancers

The assembly of reliable training samples is one of the bottlenecks in most biological applications since experiments are often expensive and/or time-consuming. In particular, labeled positive examples are rare, which is why a reliable prediction method is needed in the first place. We choose training enhancers combining two enhancer properties: the presence of bidirectional transcripts and chromatin accessibility. In the first part of the definition of our training enhancers, we chose all enhancers from the FANTOM5 database that showed a certain amount of bidirectional transcription in our cell type or tissue of interest for one or more replicates. Here, the level of transcription corresponds to the level of observed tag counts measured with the CAGE technique (see Sections 2.4.6 and 2.3.3 for more details on the FANTOM5 database and CAGE, respectively). The criteria used on the number of counts

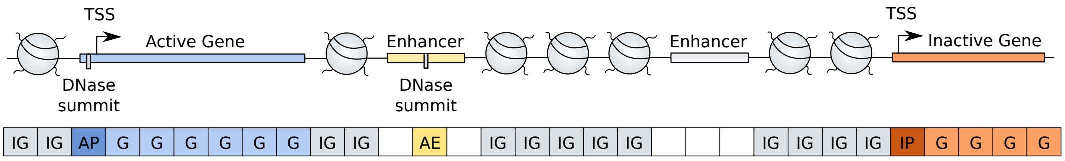


Figure 4.3: Genomic region annotation. Each 100 bp bin overlapping with a TSS of an active gene (light blue, $\log_2(\text{FPKM} + 1) > 2$) and with a DNase-seq peak (summit, light gray) is defined as an active promoter ('AP', blue), whereas bins overlapping with a TSS of an inactive gene (light orange, FPKM value = 0) are marked as inactive promoter ('IP', orange). Active enhancer bins ('AE', yellow) are mapped to FANTOM5 annotated enhancers containing a DNase-seq peak (light yellow) marking the accessible region within the enhancer. The remaining part of the active enhancer as well as FANTOM5 enhancers without corresponding DNase-seq peak are not used for training (white bins). All bins which are not overlapping with enhancers, promoters or genes are assigned to the group of intergenic regions ('IG', gray).

can vary for different cell types to adjust the amount of resulting enhancers and is summarized in Table B1.

In a second step, we discard all identified FANTOM5-based enhancers from our training set which do not overlap with a DNase-seq peak (peak calling details in Ramisch *et al.* (2018)) and as such are not accessible in the corresponding cell type or tissue. This is necessary to filter our training set for possible false positives stemming from CAGE experiments which may not exactly match the experimental setting of the HM ChIP-seq data used for the construction of the feature set. The DNase-seq experiments have been done under the same conditions as the HM data and as such constitute a reliable filter. Finally, every 100 bp bin overlapping a FANTOM5 annotated enhancer and a DNase-seq peak is defined as an active enhancer for training and testing, which is also depicted in Figure 4.3.

As a result, the active enhancers are centered on their accessible region, which is crucial for the successful training of our model. The final number of active enhancers per cell type can be found in the Appendix, Table B2.

4.1.6 Definition of training promoters

In the training set of both classifiers, but especially in the one of classifier 2, we need to distinguish active from inactive promoters. We make this decision based on a combination of accessibility and gene expression data since, following the same line of thought as for enhancers in Section 4.1.5, the criterion should be as feature-independent as possible. We could not use CAGE data for our promoter definition, as there was no data available for TSSs in our cell types of interest.

We compute FPKM (fragments per kilobase million) gene expression values from RNA-seq data matching the cell type of our HM data origin. The FPKM values were computed with *DESeq2* (R package version 1.20.0, Love *et al.* (2014)). Then, we define annotated genes from the Ensembl database (human assembly 'GRCh37.70', mouse assembly 'GRCm38.90', Kersey *et al.* (2018)) as active if $\log_2(\text{FPKM}+1) > 2$, and as inactive if $\log_2(\text{FPKM}+1) = 0$. In case that multiple replicates are available, this has to hold for all of them accordingly. Our (strict) cutoff is based on the expression profile of $\log_2$ transformed FPKM values depicted in Figure 4.4. Note also that, even though $\text{FPKM} = 1$ (equivalent to $\log_2(\text{FPKM}+1) = 1$) is often used, a widely excepted generic cutoff to distinguish expressed from not expressed genes does not exist.

We define each 100 bp bin overlapping an annotated TSS of an inactive gene as inactive promoter. For the definition of active promoters, we also incorporated the accessibility represented by DNase-seq data to center the promoter on its open region. Active promoters are therefore defined as 100 bp bins overlapping with an annotated TSS of an active gene and a DNase-seq peak. These concepts are also illustrated in Figure 4.3 and an overview of the final number of active and inactive promoters in the different cell types and tissues can be found in Table B3. TSSs of genes with $0 < \log_2(\text{FPKM}+1) \leq 2$ were not used for training.

4.1.7 Enhancer annotation

To call enhancers in a genome-wide manner, we make enhancer predictions by multiplying the outcome of classifier 1 and classifier 2 (introduced in Sec-

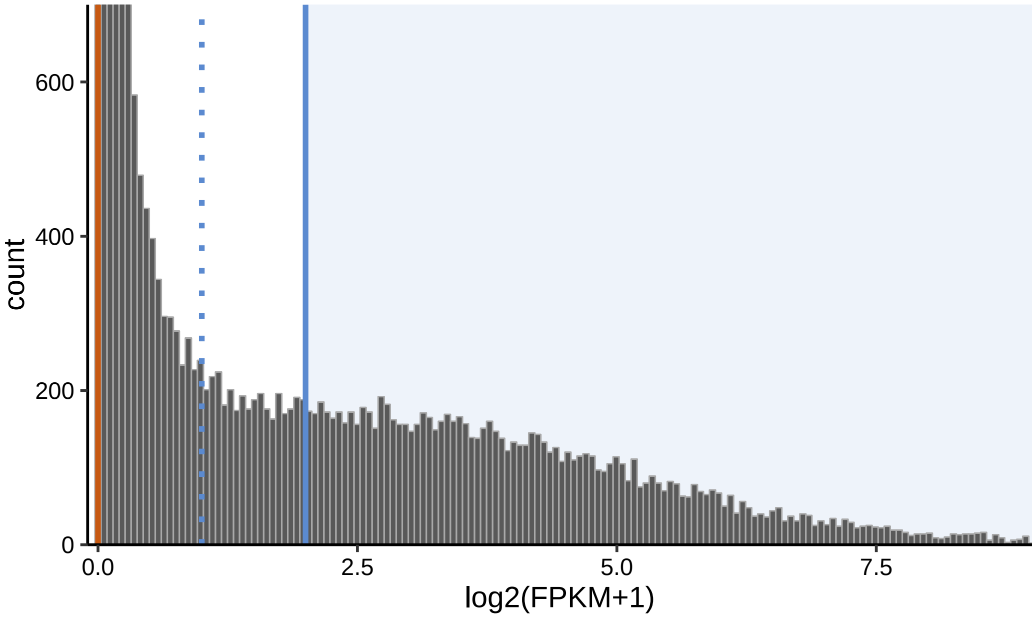


Figure 4.4: Example distribution of gene expression values. Distribution of FPKM normalized gene expression values per gene for a mouse embryonic stem cell RNA-seq sample. We defined a gene with $\log_2(\text{FPKM} + 1) > 2$ as active (blue solid line), and every gene with $\log_2(\text{FPKM} + 1) = 0$ as inactive (orange solid line). The dotted blue line indicates a weaker cutoff, that we will come back to in Section 4.2.2.

tion 4.1.2) for each 100 bp bin in the genome. Then, we sort all bins with an enhancer probability of ≥ 0.5 in a descending order from highest to lowest probability. Bins with identical probabilities are sorted according to their genomic location per chromosome where bins closer to the start of the chromosome come first/higher in the sorted list. In a next step, we extend all bins to 1100 bp by adding 500 bp upstream and downstream. This is motivated by considering $N = 5$ neighboring bins to the right and to the left in the feature set (see also Figure 4.2). Starting at the most probable enhancer bin, we detect all overlapping bins and discard them from the sorted list. We repeat this procedure for each bin in the list moving from highest to lowest probabilities. This results in a final list of non-overlapping 1100 bp enhancers with probabilities ≥ 0.5 (see Figure 4.1).

4.1.8 Regions of high enhancer density

We define regions of high enhancer density by combining proximal enhancers to bigger clusters. Following the definition of super-enhancers as stated by Hnisz *et al.* (2013) and Whyte *et al.* (2013), each pair of called enhancer peaks with a maximum distance of 12.5 kb is concatenated (see also Figure 4.1). In contrast to the definition of super-enhancers, our enhancer clusters are built solely based on distance, and do not take measures of enhancer activity or levels of TF binding into account. See also Section 2.4.9 for more details on super-enhancers.

4.2 Assessment of training enhancers in mouse embryonic stem cells

In Chapter 5, we will validate the performance of our enhancer prediction method on the example of a mouse embryonic stem cell data set. To be able to make reliable performance statements, we first assess the quality of the corresponding HM ChIP-seq samples and analyze various properties of the training enhancers which we chose based on CAGE data and accessibility.

4.2.1 ChIP-seq data quality

We applied the *plotFingerprint* tool (*deepTools*, Ramírez *et al.* (2016)) to our mESC ChIP-seq samples to assess data quality. With this tool we can measure how well a certain HM mark can be distinguished from the genomic background signal in the input or control sample. According to what we would expect based on prior knowledge we can then make a conclusion about the quality of the individual marks. For the histone modifications H3K27ac or H3K4me3, for example, we expect a very strong and specific enrichment according to genome-wide coverage results from (Kellis *et al.*, 2014). Moreover, both marks were observed to cover only a small fraction ($< 20\%$) of the genome. H3K4me1 shows a broader and overall lower enrichment and is distributed over a third of the genome. The enrichment of transcript elongation mark H3K36me3 is

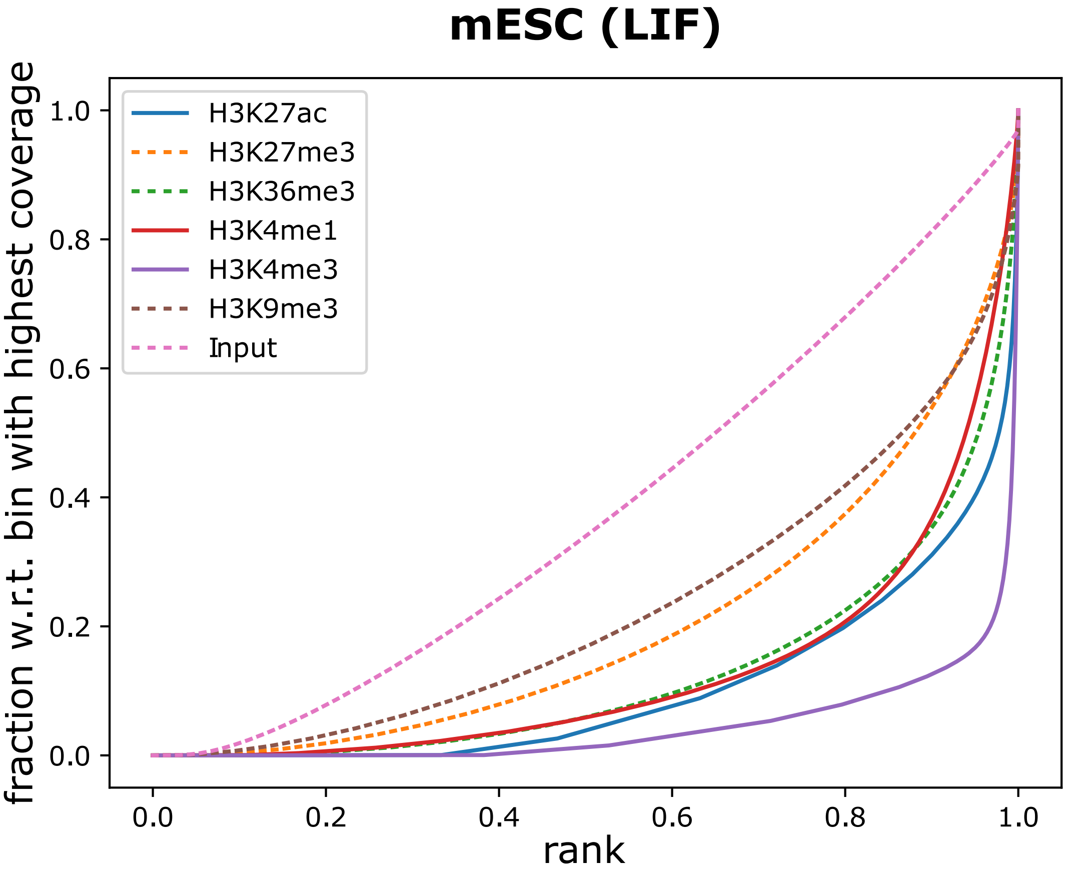


Figure 4.5: Fingerprint quality control metrics for mESC ChIP-seq experiments. For each HM ChIP-seq data, reads with a mapping quality ≥ 30 are counted per adjacent 500 bp bin. Then, the read counts are sorted and their cumulative sum is plotted. The plots are done with *deepTools* (Ramírez *et al.*, 2016).

also very broad with more than half of the enriched regions having a very low signal strength. The repressive marks H3K27me3 and H3K9me3 were observed to cover more than half of the genome with an almost exclusively very low signal.

In a first step, the genome is divided into non-overlapping 500 bp bins for which reads with a mapping quality ≥ 30 are counted. Then, the computed read counts are sorted from lowest to highest and their cumulative sum is plotted. This is done for each sample separately.

For the perfect control or input sample, we expect a profile close to the diagonal indicating a uniform and hence unspecific read distribution throughout the genome. In Figure 4.5, we see that the input sample for the mESC data set fulfills this criteria. The profile of H3K4me3 shows a steep rise for the highest ranked genomic bins, likely a result of specific enrichment which should be well differentiable from the background. More specifically, the cumulative sum over 95% of the genomic bins covers less than 20% of the highest read coverage. Furthermore, we can see the proportion of the genome which was not sequenced at all. For H3K4me3, nearly 40% of the genomic bins are not covered by any reads which coincides with the observations from Kellis *et al.* (2014). The other HM samples also fulfill the expectations according to genome coverage and signal strength distribution. For the repressive marks, nearly all genomic bins are covered with reads, but mostly at a very low number. H3K27ac, H3K4me1 and H3K36me3 behave similarly. However, H3K27ac has a higher number of bins without read coverage indicating a more specific enrichment.

4.2.2 Properties of FANTOM5 training enhancers

According to Section 4.1.5, we define active enhancers for training based on their accessibility as well as on their ability to produce transcripts in a bidirectional fashion, which can be measured with the CAGE technique (see Section 2.3.3). The FANTOM5 enhancer data base for mouse (genome build ‘GRCm37’) consists of $44,459$ potential enhancer regions and the corresponding CAGE counts in $1,037$ samples of different cell lines including biological replicates. For human enhancers (genome build ‘GRCh37’), CAGE counts for $65,423$ potential enhancer regions in $1,828$ samples was collected.

In total, the FANTOM5 data base covers $1,037$ samples covering many different cell types and tissues. We define candidate enhancers for our mouse embryonic stem cell (mESC) data based on a subset of 13 samples:

- ES-OS25 embryonic stem cells, DMSO control – 3 biological replicates
- ES-OS25 embryonic stem cells, untreated control – 3 biological replicates
- ES-Ert embryonic stem cells, untreated control, 48hr – 3 biological replicates

- ES-OS25 embryonic stem cells, untreated siRNA control – 2 biological replicates
- ES-OS25 embryonic stem cells, scrambled siRNA control – 2 biological replicates

From the $44,459$ putative FANTOM5 enhancers, we found 372 with ≥ 4 CAGE counts in all 13 samples. Even though a count cutoff of 3 is denoted as a 'permissive' cutoff for promoters, our chosen filter should lead to confident enhancer candidates since Andersson *et al.* (2014) found that enhancers have an overall lower RNA abundance than promoters. We converted the 'GRCm37' genome coordinates to 'GRCm38' using the coordinate conversion tool *liftOver* from the UCSC Genome Browser utilities to match the genome build of our HM and DNase-seq data (Hinrichs *et al.*, 2006).
In total, 280 of the 372 selected FANTOM5 enhancers overlap with a DNase-seq peak in mESC and are therefore also accessible.

Length distribution of FANTOM5 enhancers

The length of a putative enhancer defined by the FANTOM5 consortium represents the merged region of transcription initiation on the minus and plus strand of the bidirectional transcripts, and as such covers the maximum length of the accessible chromatin to which the transcription machinery binds. The majority of the 280 accessible FANTOM5 enhancers have a length between 100 bp and 400 bp, as can be seen in Figure 4.6, with a minimum length of 61 bp and a maximum of 1640 bp.We concluded to base our enhancer prediction on feature information from regions up to ~ 2000 bp in length.

Distribution of CAGE tags at FANTOM5 enhancers

As described above, we chose mESC-specific enhancers by (i) filtering for FANTOM5 enhancers which have at least 4 CAGE counts in each of the 13 mESC-associated samples, and (ii) discard those regions that do not overlap with a mESC-specific DNase-seq peak. The CAGE counts in the remaining 280 enhancers over all 13 samples range from 4 to 765, which is depicted in Figure 4.7

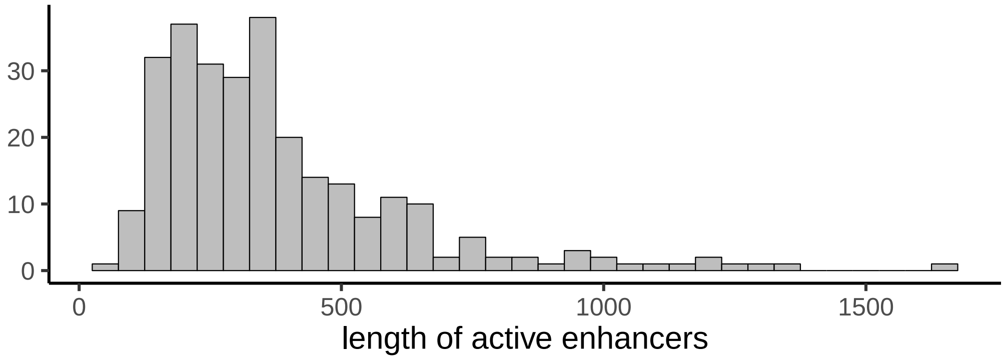


Figure 4.6: Length distribution of filtered FANTOM5 enhancer in mESC. Majority of 280 FANTOM5 enhancers found to be active in mESC and overlapping with a DNase-seq peak have a length between 100 and 400 bp. The length of the enhancers is based on CAGE data and was computed by the FANTOM5 consortium.

a). Looking at each sample individually (Figure 4.7 b) on a smaller scale, the count distributions vary even within the same cell type, but most enhancers have between 10 and 40 counts.

The FANTOM5 consortium also computed tags per million (TPM) mapped reads, which were normalized between the individual samples (see Andersson *et al.* (2014) for details). But since they did not make direct recommendations how to choose a TPM-based cutoff to define enhancers, we use a tag count oriented definition in this work. Also, we observed that it does not influence our final set of training enhancers much. Choosing a TPM cutoff of 0.5 for all 13 samples and discarding all enhancers not overlapping with a DNase-seq peak, we got 354 putative enhancer regions in mESC. Of these, 267 overlap with the 280 putative enhancers that we found previously by applying a CAGE tag count based cutoff. The TPM distributions in the 354 enhancers in all 13 samples are shown in Figure 4.8.

Histone modification patterns and accessibility at FANTOM5 enhancers

To check whether the HM profiles and the chromatin accessibility at the CAGE tag-based active enhancers display the enhancer-typical features, we computed

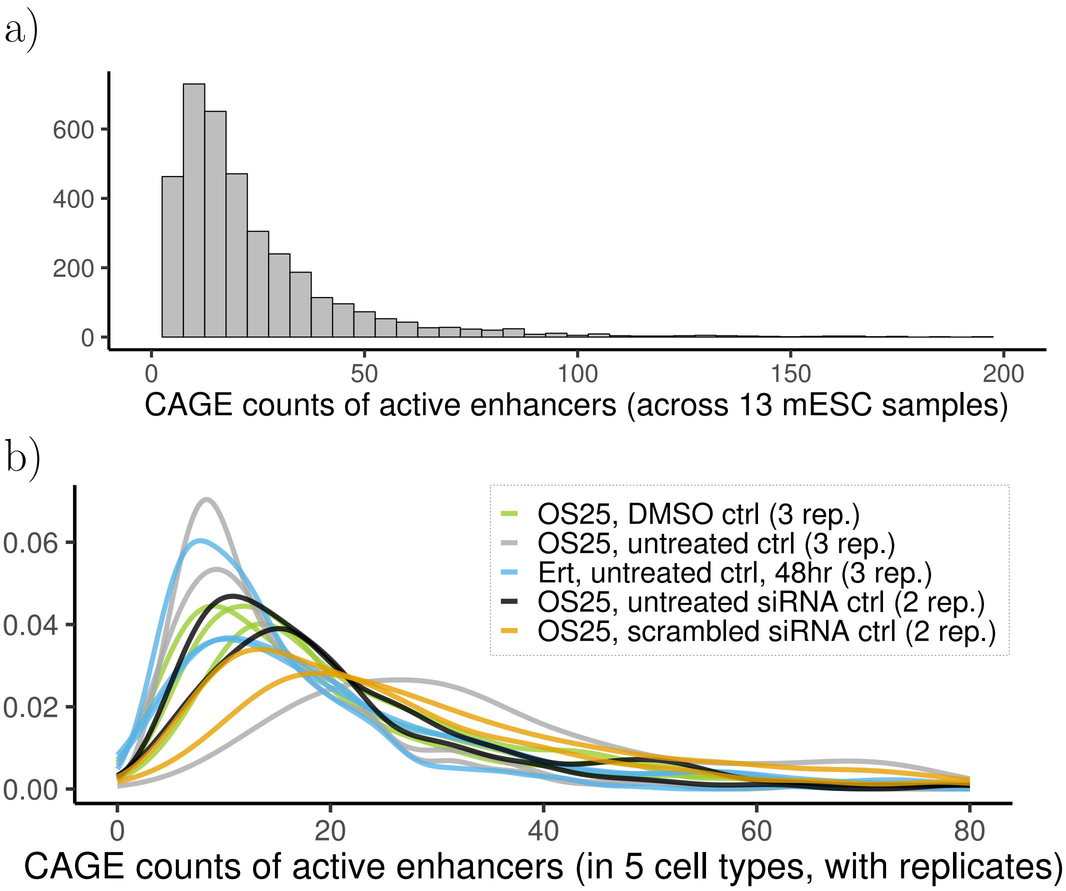


Figure 4.7: CAGE count distribution of mESC enhancer. **a)** Histogram of raw CAGE tags for 280 FANTOM5 enhancers over all 13 mESC-associated FANTOM5 samples. The x-axis is cut at 80. **b)** Density plot of raw CAGE tags for 280 FANTOM5 enhancers in individual mESC-associated samples. Identical colours indicate replicates of the same cell type. The x-axis is cut at 200.

the read counts (in 100 bp bins) in a 3 kb window around our 280 training enhancers for our six core HMs and an independent ATAC-seq data in mESC. Our filtered enhancers show a centered ATAC-seq peak indicating accessibility, which is in agreement with the filter of overlapping DNase-seq peaks (see Figure 4.9, and also Figures A1 and A2). Both enhancer-associated HM marks, H3K27ac and H3K4me1, show peaks adjacent to the accessible region. H3K27me3 is not present or even exhibits a depletion at our chosen enhancers. Interestingly, also H3K4me3, the promoter-associated HM mark, is enriched

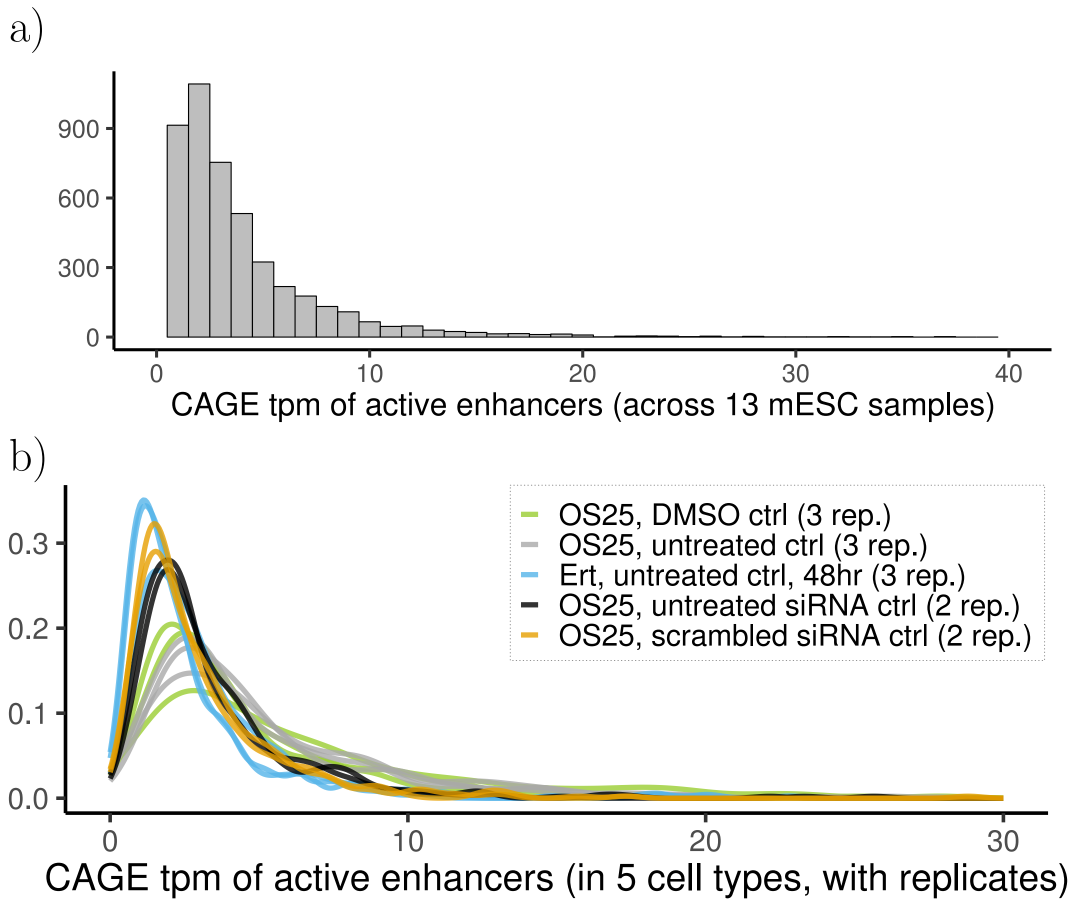


Figure 4.8: CAGE tpm distribution of mESC enhancer. **a)** Histogram of CAGE tpm values for 280 FANTOM5 enhancers over all 13 mESC-associated FANTOM5 samples. The x-axis is cut at 40. **b)** Density plot of CAGE tpm values for 280 FANTOM5 enhancers in individual mESC-associated samples. Identical colours indicate replicates of the same cell type. The x-axis is cut at 30.

to a similar degree as H3K4me1 pointing out the difficulty of distinguishing enhancers from promoters.

Histone modification patterns and accessibility at promoters

We assess the differences between enhancers and promoters in terms of HM signals by plotting the same distributions as above for our training promoters ('very active' in Figure 4.9). The promoters show a lower H3K4me1 level, and a much higher enrichment in H3K4me3. Subtle differences between the

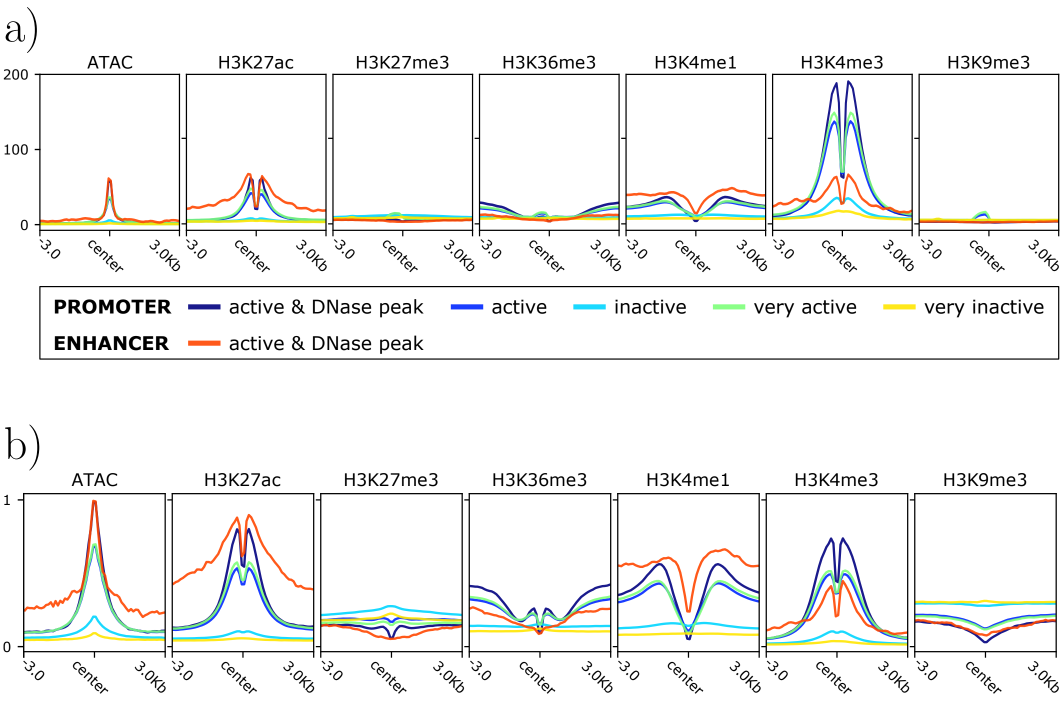


Figure 4.9: HM and ATAC-seq signal at enhancers and promoters in mESC. Summarized profiles of ATAC-seq and six HM data for six sets of genomic regions: active promoters ($\log_2(\text{FPKM}+1) \geq 1$, dark blue), 'very' active promoters ($\log_2(\text{FPKM}+1) \geq 2$, green), active promoters overlapping with a DNase-seq peak ($\log_2(\text{FPKM}+1) \geq 1$, dark blue), inactive promoters ($\log_2(\text{FPKM}+1) < 1$, light blue), 'very' inactive promoters ($\log_2(\text{FPKM}+1) = 0$, yellow), FANTOM5-based active enhancers with DNase-seq peak overlap (red). **a)** Raw cage counts. **b)** Raw cage counts excluding 1st and 95th quantile, and scaled to $[0, 1]$ for each data individually. Plots based on results from *deepTools* (Ramírez *et al.*, 2016).

different sets become especially clear in the normalized profiles excluding few regions with particularly low or high values (see Figure 4.9 b)).

Since we chose a quite strict cutoff on the expression values of the corresponding genes for our promoter definition (see Section 4.1.6, we wanted to exclude a possible bias towards a very high activity level and hence defined a second set of promoters with a less stringent cutoff. Based on Figure 4.4, we defined all promoter regions belonging to a gene with $\log_2(\text{FPKM}+1) \geq 1$ as active promoters. In Figure 4.9 it can be seen that the set of less stringently defined promoters shows on average very similar profiles compared to the more stringent set. Hence, our strict cutoff for the training promoters should not influence the learned classification rules to distinguish enhancers from promoters. Also, we seem to decrease the amount of promoters that are not active according to their HM levels by applying the stricter cutoff (see Figure A2). In our final set for training, we take the overlap of the strictly defined promoters and mESC-specific DNase-seq peaks, resulting in regions with an on average even higher H3K27ac and H3Kme3 signal. Since the promoter activity was defined based on a strict cutoff on the gene expression, inaccessible promoter regions could be due to wrongly annotated TSSs. These then decreased the signal of promoter-associated HMs on average.

4.3 Hyperparameter optimization using grid search

After ensuring the suitability of the feature data and the training regions in the previous section on the example of mouse embryonic stem cell data, we can finally train our two random forest classifiers. Both classifiers have two free hyperparameters, the number M of decision trees used in the random forests (see Section 4.1.2) and the number N of pairs of neighboring 100 bp bins (one on each side of the bin of interest) taken into account in the feature space (see Section 4.1.1), which have to be set before the training process starts.
We optimize these parameters by performing a grid search for
$M = \{10, 20, \ldots, 100\}$ and $N = \{0, 1, \ldots, 10\}$. Moreover, we sample 10 different training sets according to the rules described in Section 4.1.3 for each parameter combination (M, N) to decouple the performance results from the random factor in the training set choice. This results in $10 \cdot 11 \cdot 10 = 1100$ combinations of M, N and the training set, for which we learn classifier 1 and classifier 2, and compute the area under the precision recall curve (AUC-PR) and the area under the ROC curve (AUC-ROC) on a test set (see Section 3.3.4 for more details on these performance measures). For classifier 1, we concentrate more on the AUC-PR results, since the class of interest ('active') is underrepresented. However, for classifier 2 this is not the case and the

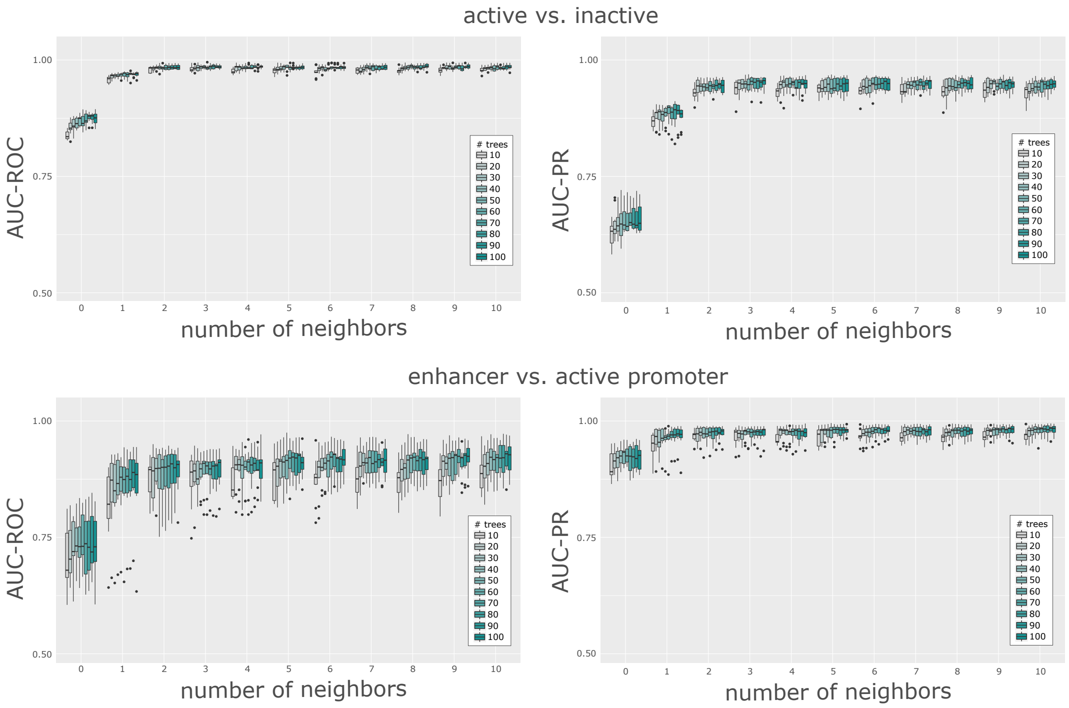


Figure 4.10: Grid search results for random forest classifiers. AUC-ROC and AUC-PR results from a 5-fold cross-validation for classifier 1 (active vs. inactive) and classifier 2 (enhancer vs. active promoter). Number of adjacent bins N included in the feature set on x-axis from 0 to 10. Different colours indicate the number of decision trees M used in the random forest classifiers.

AUC-ROC is more informative.

The performance results of both classifiers greatly depend on the number of features represented by N (see Figure 4.10). The highest change in performance happens between the choice of $N = 0$ and $N = 1$. Taking no adjacent bin into account in the learning process, i.e., predicting based on the HM values in the central 100 bp bin, leads to an AUC-PR $\in [0.58, 0.72]$ for classifier 1 and an AUC-ROC $\in [0.61, 0.85]$ for classifier 2, depending on the choice of M (see Figure 4.10). Adding HM information in one bin upstream and downstream ($N = 1$) improves the performance results to AUC-PR $\in [0.82, 0.91]$ and AUC-ROC $\in [0.63, 0.94]$, respectively.

A likely explanation is that for most active promoters and enhancers, the accessible region is at least ~ 100 bp long and therefore only little information

regarding HMs can be found in the central bin. The adjacent bins, however, seem to cover nucleosomes with modified histone tails that can help to distinguish different genomic regions. From $N = 1$ to $N = 2$ an increase in performance with a smaller range for different values of M is visible (AUC-PR $\in [0.9, 0.96]$ for classifier 1, and AUC-ROC $\in [0.75, 0.95]$ for classifier 2). Then, the performance stabilizes and only increases in small steps. We finally chose $N = 5$ for both classifier 1 and classifier 2, and thus cover $1,100$ bp of HM information for each enhancer prediction.
With increasing the number of trees M in the random forests the median performance increases independently of the number of adjacent bins N. For the rest of our analysis, we set $M = 70$ for classifier 1 and classifier 2, as a compromise between computing cost and performance.

4.4 Motif analysis

As part of the validation of our classifier, we compute the enrichment of TF motifs in different sets of predicted enhancers. Motif enrichment analysis is usually based on position frequency matrices (PFMs) which describe the binding specificity of TFs by determining the nucleotide frequencies at each position from several (aligned) TFBSs. Each PFM can be represented by a so-called motif logo as shown in in Figure 4.11. A sequence of DNA is scanned for matches with a PFM to detect the number of motif hits as well as the subsequent motif enrichment value. However, assessing the statistical significance of motif hits is a difficult task, since most motifs are rather short (a few base pairs), can include small variations in their underlying sequence without loosing their function, or lack specificity and therefore may occur by chance in the genome (A survey of motif discovery methods in an integrated framework; Geir Kjetil Sandve and Finn Drablos).
Here, we use the function *motifEnrichment* which is part of the R package *motifcounter* (Kopp, 2017; Kopp and Vingron, 2017) and was shown to outcompete various motif count models for most of the analyzed motif structures (e.g. non-self-overlapping motifs, but especially self-overlapping motifs such as

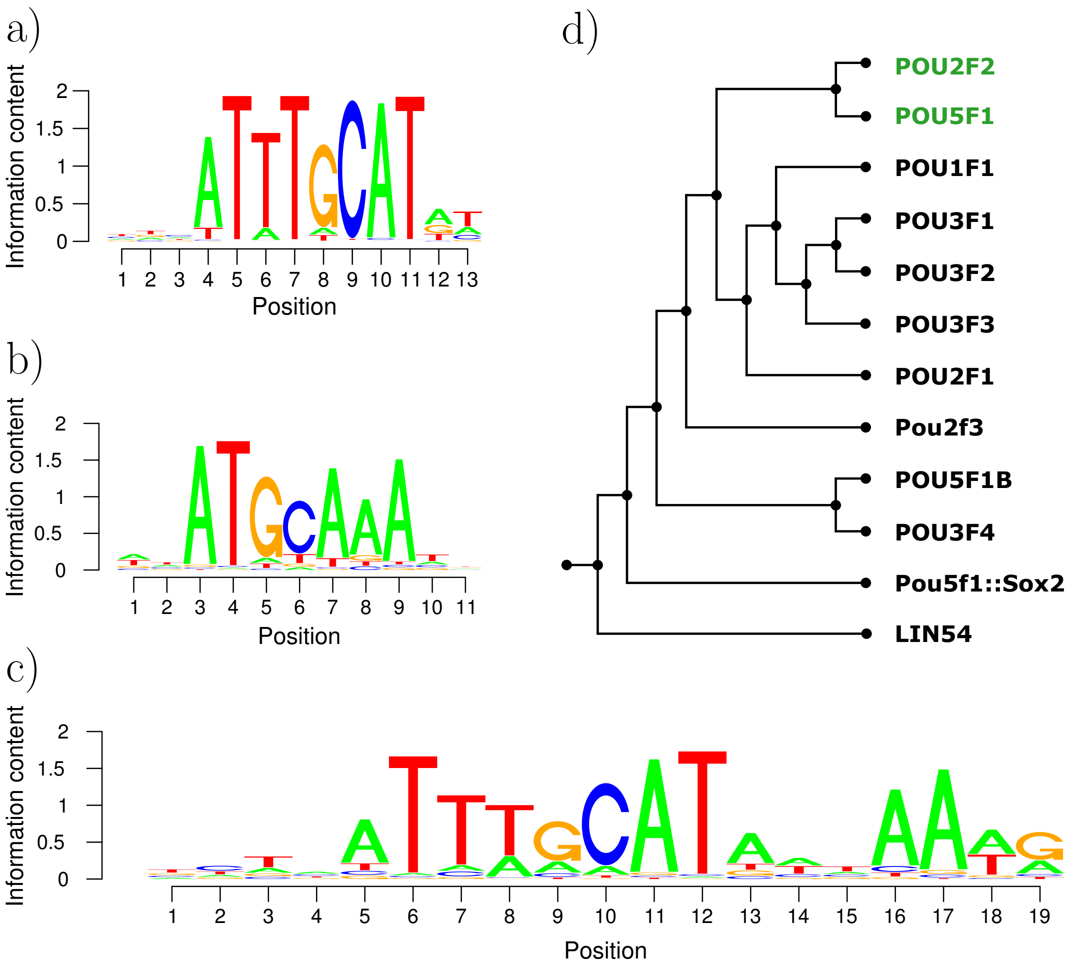


Figure 4.11: Examples of JASPAR motif logos. Motif logos of **a)** POU2F2, **b)** POU5F1 (OCT4) and **c)** the consensus motif of *cluster 18* according to the PFMs from the JASPAR CORE vertebrates database 2018. Motif logos were plotted with the R package *seqLogo* (Bembom, 2018) after normalizing the PFMs such that the columns sum up to 1. **d)** Logo tree of a motif cluster, *cluster 18*, adapted from JASPAR (Khan *et al.*, 2018).

palindromes and repeat-like motifs). It is based on a compound Poisson model and a higher-order Markov background model. Motif hits are called based on a desired false positive level, and the enrichment of motif hits is tested by comparing the number of motif hits observed in a sequence of interest with the number of motif hits in a sequence generated by the background model. By taking into account a background of higher order, possible sequence biases such as a higher frequency of 'GC' dinucleotides in CpG islands of promoters, are less likely reflected in the enrichment results.

In our analysis, we use the default parameters, an order-1 background model and a false positive level of $\alpha = 0.001$, and we scan both strands for motif hits. Also, we indicate the over-representation of a motif in a set of sequences (in comparison to the modeled background) with the fold-enrichment value.

We estimate enrichment for motifs (available as PFMs) from the open-source database JASPAR (Khan *et al.*, 2018). We use the JASPAR CORE vertebrates database which contains a set of 579 manually curated non-redundant TF motifs. In addition, to limit redundancy due to sequence similarity, the 579 TF binding motifs were clustered by tree partitioning into 78 sets (Castro-Mondragon *et al.*, 2017). An example tree for *cluster_18* is depicted in Figure 4.11 d).

Each cluster is represented by a consensus motif and corresponding logo (Figure 4.11 c)) computed from the original PFMs of the comprised TF motifs (Figure 4.11 a) and b)). We apply *motifcounter* to the consensus PFMs which we normalized such that the columns sum up to one.

4.5 Two other enhancer prediction methods

In this section, we give a detailed summary of one unsupervised and one supervised method used for enhancer prediction. The final comparison to our approach can be found in Section 5.3.

4.5.1 ChromHMM

ChromHMM is a genome segmentation tool aimed at dividing the genome into biologically meaningful combinations of chromatin marks and discovering de novo chromatin states Ernst and Kellis (2012, 2017). It is comparable to our method in terms of usability, since the necessary biological input data are ChIP-seq bam files, for example from several HMs. ChromHMM is an unsupervised multivariate HMM and as such models multiple input measurements as the observable output generated by a fixed number of hidden states, here the underlying chromatin states. Typical for HMM models, the only input parameter is the number of hidden states or chromatin states that has to be chosen by the user. The observable output is based on a set of binarized chromatin marks, where either the presence of a mark (1) or the absence (0) is encoded at each genomic position. As a result, the emission probabilities in each hidden state are modeled as a product of independent Bernoulli variables. The emission and also the state transition parameters are learned in an unsupervised manner using a variant of the iterative EM based Baum-Welch algorithm.

Following the notation in Ernst and Kellis (2010), in a first step, each chromosome $c = \{1, \dots, C\}$ in the genome is split into T_c non-overlapping 200 bp bins c_t, $t = 1, \dots, T_c$. Then, for each bin c_t, the presence of a chromatin mark $m = 1, \dots, M$ is denoted as $\nu_{c_t,m} = 1$ and the absence as $\nu_{c_t,m} = 0$. The vector $\nu_{c_t} = (\nu_{c_t,1}, \dots, \nu_{c_t,M})$ denotes the presence or absence of all M chromatin marks in bin c_t. Furthermore, let $p_{k,m}$ be the emission probability that chromatin mark m is present int state k, b_i the transition probability from state i to state j with $i, j \in \{1, \dots, K\}$, and a_i the initial probability that the first bin of a chromosome is in state i. The joint probability of a hidden state sequence $s_c = (s_{c_1}, \dots, s_{c_{T_c}}) \in S_c$ throughout chromosome c and an emitted output sequence of chromatin marks $\nu_c = (\nu_{c_1}, \dots, \nu_{c_{T_c}})$, can be written as

$$\mathbb{P}(\nu_c, s_c) = \underbrace{a_{s_{c_1}} \left(\prod_{m=1}^{M} p_{s_{c_1},m}^{\nu_{c_1,m}} \left(1 - p_{s_{c_1},m}\right)^{1-\nu_{c_1,m}} \right)}_{\text{initial state and emission probability}} \cdot$$

$$\prod_{t=2}^{T_c} b_{s_{c_{t-1}},s_{c_t}} \left(\prod_{m=1}^{M} p_{s_{c_t},m}^{\nu_{c_t,m}} \left(1 - p_{s_{c_t},m}\right)^{1-\nu_{c_t,m}} \right).$$

The unknown parameters can be estimated by maximizing the corresponding likelihood function of the genome-wide observed chromatin profiles,

$$\mathbb{P}(\nu; a, b, p) = \prod_{c \in C} \sum_{s_c \in S_c} \mathbb{P}(\nu_c, s_c),$$

for a fixed number of states K using a (computationally less time consuming) variant of the EM-based Baum-Welch algorithm (see Section 3.5.3 for more

details on HMMs and (Ernst and Kellis, 2012) for a detailed explanation of the model learning step).
Each genomic bin is finally associated with one of the K 'hidden' states based on the posterior probability distribution computed with a forward-backward algorithm.

4.5.2 REPTILE

REPTILE, short for regulatory element prediction based on tissue-specific local epigenetic marks, is a random forest based supervised method to identify enhancers from HM ChIP-seq and whole-genome cytosine DNA methylation profiles (He *et al.*, 2017). The putative enhancers used for training are promoter distal p300 binding sites.
The classification model comprises two individually learned random forests, where both consist of $2,000$ decision trees and are based on the same feature set. In total, 14 different features are used. For each region in the training or test set, DNA methylation and six HM signals (H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K27ac and H3K9ac) are computed in the target sample, i.e., the sample in which enhancer predictions are needed. Additionally, these seven features are also computed in (several) reference samples from different cell types tissues. The mean of the resulting feature levels over the reference samples is subsequently subtracted from the one in the target sample creating the so-called intensity deviation feature set, which is supposed to measure the variation between different cell types or tissues. The classifiers differ in the composition of their training sets including differently seized query regions and different strategies for labeling. One classifier is trained on $40,000$ regions of size 2 kb, of which $5,000$ are labeled as active enhancers and the rest as inactive. The putative enhancers are the top $5,000$ (promoter-distal) binding sites of the histone acetyltransferase *p300*, and the $35,000$ non-enhancer regions are composed of $5,000$ annotated GENCODE promoters and $30,000$ random genomic regions. The training set of the second classifier is based on differentially methylated regions (DMRs), which are called from the DNA methylation data in the target sample in a genome-wide manner. DMRs over-

lapping with putative enhancers from the training set of the first classifier are part of the positive set, while those overlapping previously annotated non-enhancer regions build the negative set for training.

After both classifiers are trained, the enhancer probability of a new region (default: 2 kb) is determined by taking the maximum of the enhancer score from the first classifier for the entire region and the enhancer score(s) from the second classifier for the DMR(s) within this region. Genome-wide score predictions are computed for 2 kb bins which are sliding in steps of 100 bp, as well as for all DMRs. Every DMR with an enhancer score > 0.5 is annotated as 'enhancer-like' DMR, and predictions for the 2 kb bins are annotated like described for our method (see Section 4.1.7). The final list of called enhancers is the union of both annotation results.

REPTILE is offered as a pre-trained classifier for mouse, trained on mESC data, in several settings. In contrast to our method (and also to ChromHMM), the user has to run several independent scripts for training and also for enhancer predictions. Unfortunately, the normalization is not part of the REPTILE tool and has to be done by the user beforehand. The recommendations for normalizing the ChIP-seq HM data prior to training and predicting is taking the $\log_2$ fold change relative to a control sample of the RPM (read per million) count values.

4.6 Summary

In this chapter, we introduced a new supervised classification model to predict cell type-specific active enhancers genome-wide. The predictions are based on six widely-used histone modifications (HM) and therefore collapse several epigenetic marks into one enhancer probability track.

For the design of our feature set, we took into account the observed local chromatin structure of an active enhancer, which is in essence an accessible region flanked by nucleosomes with specific HM patterns. Furthermore, our model consists of two individually trained random forests which split the task of enhancer prediction. While one random forest learns to distinguish active from inactive genomic regions, the other specializes on the more difficult distinction of active enhancers and active promoters. The composition of our training sets follows the expected proportions of different genomic elements genome-wide. A limiting factor is often the availability of gold-standard enhancers in the cell type of interest, since only few enhancers have been experimentally validated. Therefore, we took advantage of known enhancer properties and defined active enhancers in a cell type-specific way based on accessibility and their ability to produce short bidirectional transcripts. Here, we paid special attention to the independence of feature and training set criteria to avoid circular reasoning.

We introduced a mouse embryonic stem cell data set, with which we illustrated certain properties of our defined training enhancers and showed the suitability of the training and the feature set for our enhancer prediction task. We optimized both random forest classifiers and observed that the length of the accessible region within an enhancer is reflected in the optimal feature choice. Finally, we described a motif analysis framework which we will apply in the following chapters for evaluation purposes, and introduced two competitor methods commonly used for enhancer prediction.

5 Validation of our Enhancer Prediction Method

In this chapter, we extensively validate our enhancer prediction method which was introduced in the previous Chapter 4. We present our prediction results when trained and tested in a single cell type, and also offer a pre-trained classifier that can be reliably applied across different cell types and species solely based on the corresponding HM data. We compare our enhancer prediction method with a commonly used unsupervised approach, ChromHMM, and a recently published supervised method, REPTILE, which is in part based on histone modification data.

The presented results are (in a less detailed manner) published as a pre-print (Ramisch *et al.*, 2018), which I co-first authored with Verena Heinrich. The content of this chapter is soley based on my part of the work, except for the application of the ChromHMM genome segmentation tool on the mESC data set in Section 5.3, which was done by Philipp Benner, a co-author of the above mentioned paper.

5.1 Results of enhancer prediction in mouse embryonic stem cells

In Section 4.3, we performed a hyperparameter optimization for our two random forest classifiers trained on a mouse embryonic stem cell (mESC) data set. The optimized classifiers are both based on $M = 70$ decision trees and $N = 5$ pairs of adjacent bins in the feature set. Below, we discuss the importance of our features for the classification decisions and subsequently present the performance results of the final enhancer prediction method. More precisely, we assess the performance for a chosen test set, measure the spatial resolution of our genome-wide predicted enhancers as well as the validity of the promoter-proximal enhancer predictions. For further validation, we compare regions with a high predicted enhancer density with already annotated super enhancers and finally perform a motif analysis on our enhancer set.

5.1.1 Feature importance of optimized random forest classifiers

During training, a random forest classifier collects information on the importance of the features. One of the importance measures is the Gini index, which we introduced in Section 3.4.3. The Gini index allows to recapitulate which features were preferentially chosen to optimize split decisions in the individual decision trees of the forest. Since at each split only a randomly drawn subset of features is tested for optimization, highly correlated features are expected to show similar results in terms of their Gini index.

For our enhancer prediction method, we analyze the feature importance of the two optimized random forest classifiers, classifier 1 (active vs. inactive regions) and classifier 2 (enhancer vs. active promoter). The feature set of classifier 1 is based on the active mark H3K27ac and the two repressive marks H3K27me3 and H3K9me3. This results in $3 \cdot 11 = 33$ features in total taking into account $N = 5$ adjacent bins upstream and downstream of the bin of interest, denoted by $\{$ $\text{bin}_{x-5}, \ldots, \text{bin}_{x-1}, \text{bin}_x, \text{bin}_{x+1}, \ldots, \text{bin}_{x+5}\}$. We can see in Figure 5.1 that H3K27ac is by far the most important feature distinguishing an active from an inactive genomic region. The central bin_x as well as $\text{bin}_{x\pm1}$ show a low Gini index, while for the second adjacent bin to both sides, $\text{bin}_{x\pm2}$, a clear increase resulting in a peak of importance is visible. This suggests that the central 300 bp bin of an active enhancer is on average accessible/ nucleosome-free and as such shows no HMs. Furthermore, the importance measure decreases in $\text{bin}_{x\pm5}$, which supports the choice of taking $N = 5$ adjacent bins to both side into account. The importance peak of H3K27ac to the left of bin_x is

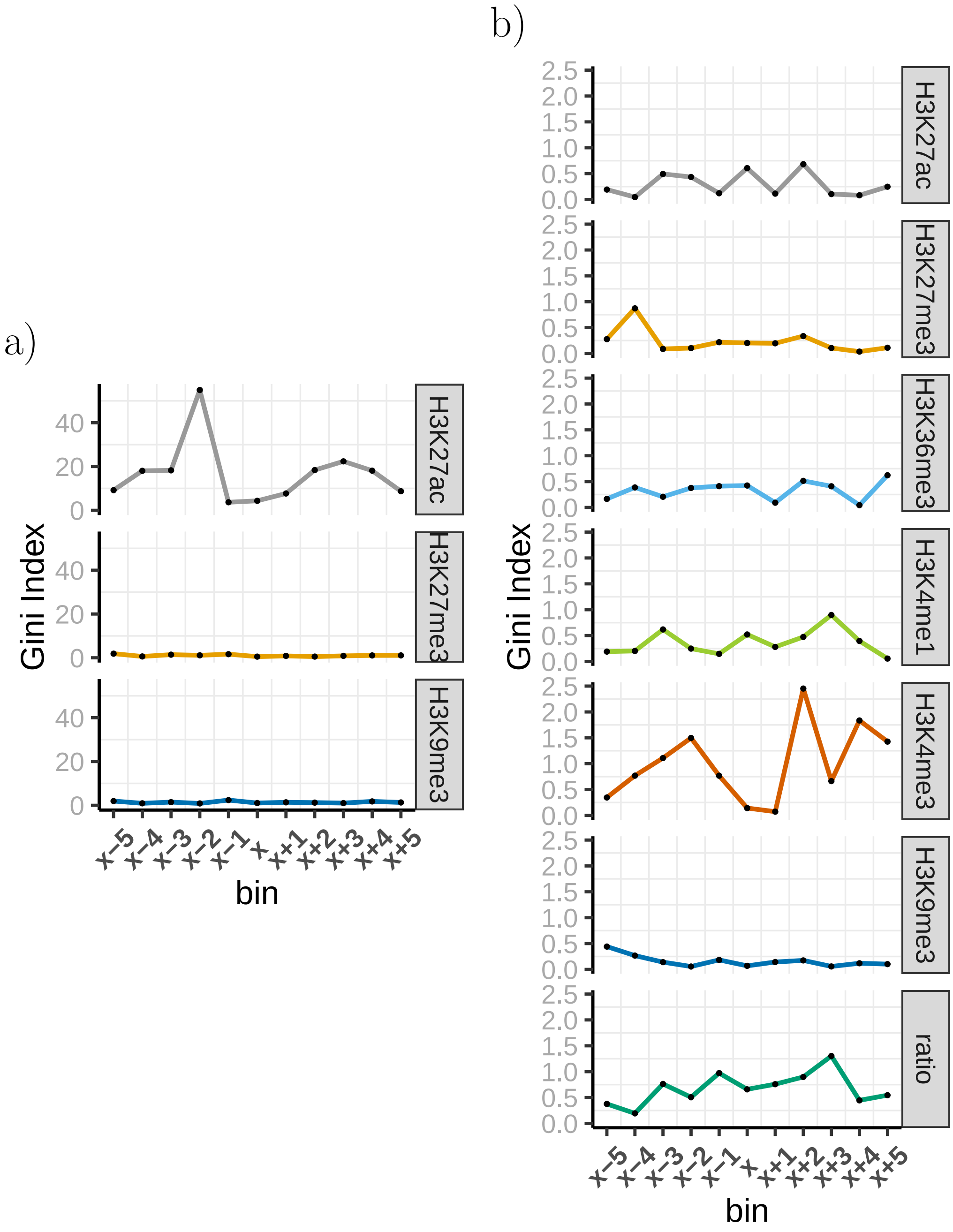


Figure 5.1: Feature importance of optimized random forests. Gini index for 11 bins per HM-based features on the y-axes and genomic bins on the x-axes. Bin x denotes the central bin of the genomic training regions. **a)** Classifier 1 (active vs. inactive) with three HMs. **b)** Classifier 2 (enhancer vs. active promoter) with six HMs and the H3K4me1/H3K4me3 ratio.

higher than on the other side which could be due to the randomness in the feature set choice at each split, i.e., bin$_{x-1}$ was chosen more often by chance. Another possibility is that the training set consists of stronger plus than minus strand promoter which usually have asymmetrical H3K27ac peak heights. We balanced the number of both types of promoters, but the choice according to activity strength is random and could have an effect on the feature importance. The classifier learning the difference between active enhancers and active promoters, classifier 2, is based on the six HMs H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3, and the ratio H3K4me1/H3K4me3. The feature with the highest importance values is H3K4me3 which is low in the three center bins and peaks at positions bin$_{x\pm2}$. This agrees with the observations in Figure 4.9, that promoters and enhancers show similar H3K4me3 level, but active promoters show a much higher H3K4me3 signal on average. The H3K4me1/ H3K4me3 ratio is the second most important feature with higher values in the seven central bins, bin$_{(x-3):(x+3)}$. H3K4me1 has overall low importance values, but still shows small peaks at bin$_{x\pm3}$, which could hint towards a broader enrichment of H3K4me1 at enhancers than at promoters (see also Figures A3 and A4).

5.1.2 Performance results for test set prediction

We validate our mESC classifier based on the optimized hyperparameters $N = 5$ and $M = 70$ on an independent test set containing 1000 genomic regions (see Section 4.1.3 for a more detailed description of the test set). First, we confirm that our chosen test set is representative of the composition of the genome in terms of its HM signature and accessibility. After that, we compute the performance results using the area under the ROC curve (AUC-ROC) and area under the precision recall curve (AUC-PR). We also discuss possible reasons for difficulties in performance validation.

Test set is representative of genome

Intergenic regions, inactive promoters and intragenic regions from active genes in our test set should be easily distinguishable from active enhancers since they

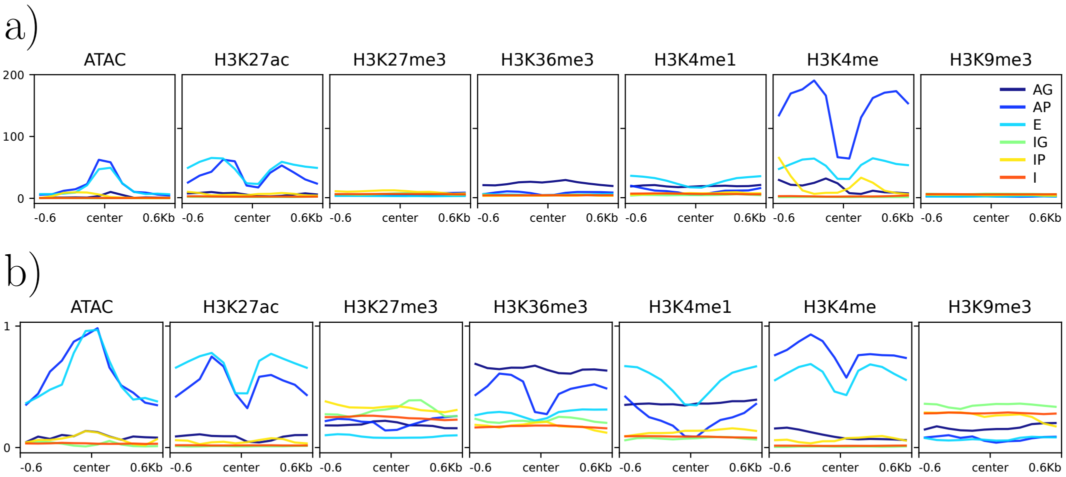


Figure 5.2: Histone modifications and accessibility at test set regions. HM and ATAC-seq profiles for all active genes (AG, dark blue), active promoters (AP, blue), active enhancers (E, light blue), inactive genes (IG, green), inactive promoters (IP, yellow) and intergenic regions (I, red) in the test set. **a)** Raw cage counts. **b)** Raw cage counts excluding 1st and 95th quantile, and scaled to $[0, 1]$ for each data set individually. Plots based on results from *deepTools* (Ramírez *et al.*, 2016).

are not accessible and do not carry H3K27ac and H3K4me1 (see Figures 5.2 and A5) but instead are slightly enriched for the repressive marks H3K27me3 and H3K9me3. Genomic regions placed in active genes are slightly enriched for H3K4me1, but in a uniform fashion, i.e., without the enhancer or promoter typical pattern of an accessible region with surrounding peaks of enrichment. Also, they do not carry the active mark H3K27ac, which should make them easily distinguishable from enhancers through classifier 1 (active vs. inactive). Enhancers and active promoters have a very similar profile regarding the accessibility and the active mark H3K27ac. On average, the test set enhancers and active promoters differ in H3K4me3 and H3K4me1 profiles, and also the active promoters show the transcriptional elongation mark H3K36me3.

Based on these observations, our test set seems to be representative regarding challenges expected for genome-wide enhancer predictions, and therefore measured test set performances should serve as a good indicator for the quality of our classifier model.

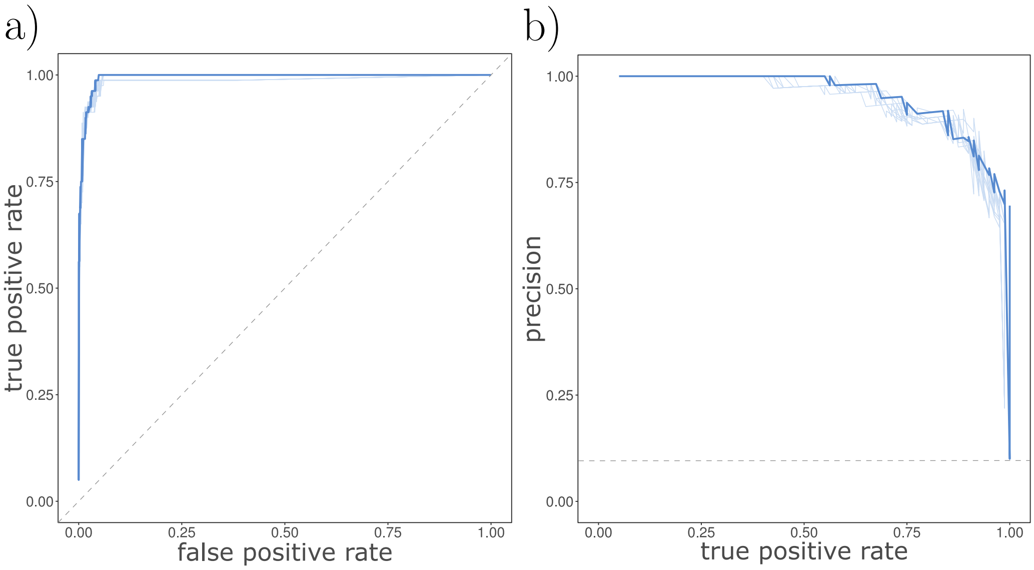


Figure 5.3: ROC and precision-recall curve for test set predictions. Performance results for ten randomly sampled (overlapping) test sets. Grey lines indicate the expected performance result for a random classifier. **a)** ROC curve. **b)** Precision-recall curve.

ROC-curve and Precision-recall curve

We measure the performance of our mESC enhancer classifier in terms of area under the receiver operating statistics curve (AUC-ROC) and the area under precision-recall curve (AUC-PR) on a randomly sampled test set described in Section 4.1.3. To control for the performance variation between different test set choices, we sampled 10 overlapping test sets containing 800 regions from the original test set (positive and negative regions are kept at the same ratio). We apply our optimized classifier to the 10 test sets and predict enhancer probabilities for 800 regions each time. Sliding the enhancer probability cutoff from 0 to 1 in steps of ~ 0.005 we compute 200 true positive rates (TPR), false positive rates (FPR) and precision values (PREC) with which we compute the AUC-ROC and AUC-PR (more details on the performance measures can be found in Section 3.3.4). Overall, the AUC-ROC as well as the AUC-PR are stable for these different subsets taking very high values in $[0.98, 0.99]$ and $[0.93, 0.96]$, respectively, as can be seen in Figure 5.3.

Choosing the test set with the best AUC-PR performance (shown in dark blue in Figure 5.3) and a cutoff of 0.5, i.e., a classification according to

$$\text{label}(\text{bin}_x) = \begin{cases} 0 \text{ (non-enhancer)}, & \text{if } \mathbb{P}(\text{bin}_x = \text{active enhancer}) < 0.5 \\ 1 \text{ (enhancer)}, & \text{if } \mathbb{P}(\text{bin}_x = \text{active enhancer}) \geq 0.5, \end{cases}$$

the results for our classifier are TPR= 0.85, FPR= 0.015 and PREC= 0.86. In words, this means that

- 85% of the true enhancers in the test set were also predicted to be enhancers (TPR),
- 86% of our predicted enhancers are also true enhancers (PREC), and
- 1.5% of the non-enhancer regions in the test set were wrongly predicted to be enhancers (FPR).

Nearly half (5/12) of the wrongly classified true enhancers, called false negatives, have a predicted enhancer probability higher than 0.43 and thus missed the cutoff to be labeled as an enhancer by less than 0.1.
From the wrongly classified non-enhancer regions, called false positives (FPs), also nearly half (5/11) have a probability below 0.6 and are therefore based on low-confidence enhancer predictions. The original label of the false positive enhancer predictions are very heterogeneous:

- 4 intergenic regions,
- 3 active promoters (with probabilities between 0.51 and 0.67),
- 1 inactive promoter and
- 3 active genes,

from which one is the FP with the highest predicted probability (0.8). The normalized HM counts at this originally labeled active gene show indeed resemblance to HM profiles at active enhancer regions: a high H3K27ac and H3K4me1 signal with two peaks on both sides of a signal depletion at the center (see Figure 5.4). A possible reason is that by coincidence an unknown active enhancer was part of the negative subset of the test set, which could distort the performance and is a good example of why the validation of enhancer prediction methods is a difficult task. Of course, this cannot be concluded with certainty since we still lack a complete understanding of the role of HMs at genomic regions.

5.1.3 Spatial resolution for genome-wide predictions

Another way of validating our mESC classifier is to measure the spatial resolution of genome-wide enhancer predictions, i.e., the distance to the closest accessible region represented by ATAC-seq data. We annotate enhancers as described in Section 4.1.7 and compute for each single enhancer the distance to the closest ATAC-seq summit. More details on the ATAC-seq data set and the peak calling can be found in Ramisch *et al.* (2018). With a probability cutoff of 0.975, we call the top $1,484$ enhancers, which have a median distance of 118 bp to the closest accessible region as can be seen in Figure 5.5. Adding $1,779$ enhancers with a predicted probability $\in [0.95, 0.975)$ to the total set of enhancers leads to a slightly increased spatial resolution which is still under 150 bp. For the top $\sim 11,000$ enhancers at a cutoff of 0.825 we get a good resolution of 236 bp, and taking together all $42,530$ enhancer with a probability ≥ 0.5 we are still within ~ 430 bp of the closest accessible region.

5.1.4 Promoter-proximal enhancer predictions

In comparison to other types of genomic regions, it is most challenging to distinguish active enhancers from active promoters due to similarities in a range of properties, for example similar levels of the active mark H3K27ac (see Section 2.4). We mostly succeeded to make the distinction in our test set, for which we are able to check for this specific type of false positive predictions, since we put a lot of effort in labeling the test regions (see Section 4.1.3). On a genome-wide scale, a typical criterion to measure this quality of enhancer predictions is to check the percentage of predicted enhancers which show an overlap with annotated active promoters.

We divide the called enhancers into subsets according to their predicted probabilities. This is done in steps of 0.05, resulting in ,e.g., $1,484$ enhancers with

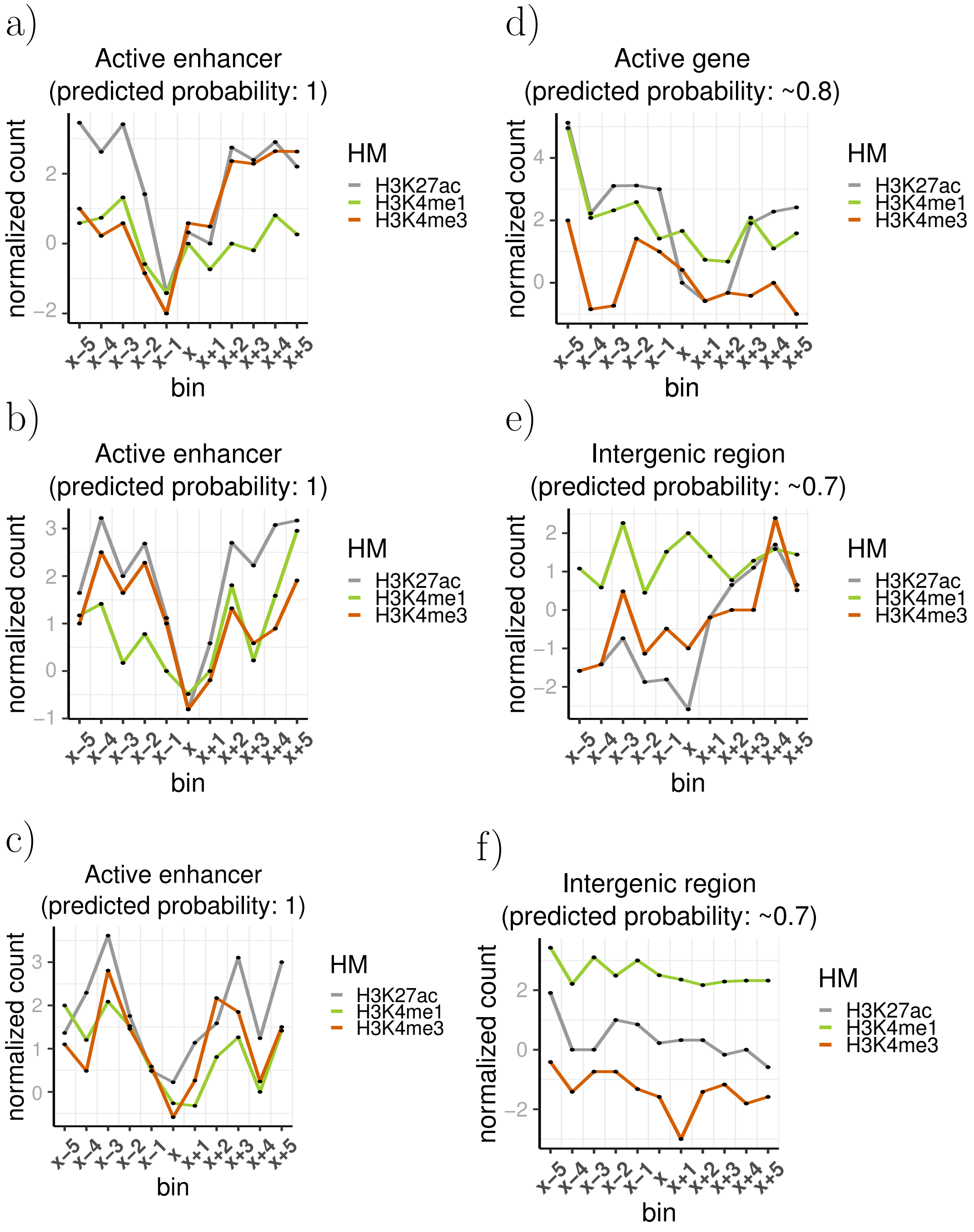


Figure 5.4: Histone modification profiles for correctly and wrongly predicted genomic regions. Input-normalized HM counts on the y-axes and genomic bin on the x-axes.

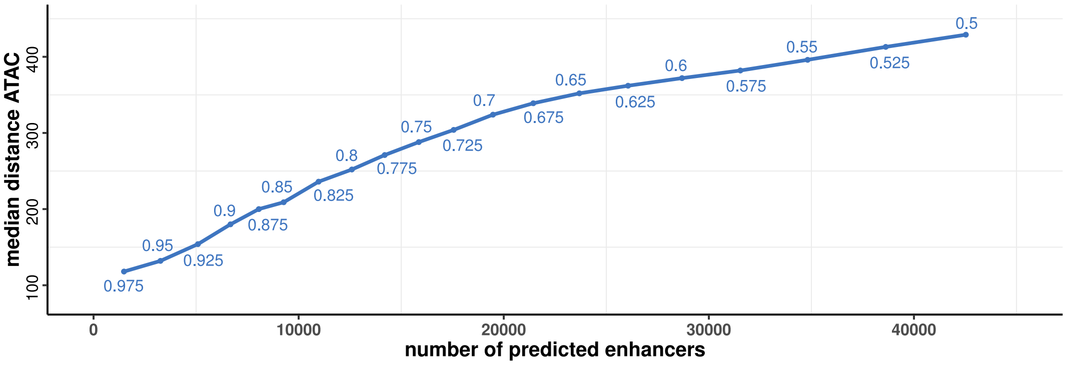


Figure 5.5: Distance to closest accessible region. We predicted enhancers genome-wide for a sliding probability cutoff from 1 to 0.5 (probabilities appear as numbers in plot). For each set of predicted enhancers, we computed the median distance to the closest ATAC-seq peak.

a probability $\in [0.975, 1]$ or 3907 enhancers with probabilities $\in [0.5, 0.525)$). For each of this subsets, we compute the distance to the nearest annotated TSS per annotated enhancer (center). Here, we use the annotation from the Ensembl mouse assembly 'GRCm38.90' and included $52,636$ TSSs promoters in total into our analysis (Kersey *et al.*, 2018). Then, we computed for each subset the fraction of enhancers falling below a certain distance threshold.

For the top ~ 5000 predicted enhancers (union of first three most confidently predicted enhancer subsets) $\sim 2\%$ are within a range of 200 bp of an annotated TSS (see Figure 5.6 a)). The low number of overlapping enhancers can also be visually confirmed by the HM profiles for the top $1,484$ enhancers in Figure 5.7 a). The percentage of overlap is highest ($\sim 9\%$) for predicted probabilities between $0.6 - 0.625$ and has a decreasing trend from there on. Since we chose a quite small distance threshold here, it is likely (but not certain) that these enhancers are indeed wrongly classified. However, taking into account the level of promoter-enhancer similarities, this is a good performance result on a genome-wide scale.

For higher distance thresholds, it is much harder to decide if overlapping enhancers are just promoter-proximal or actual false positive (FP) predictions. To pursue this question, we repeated the same analysis for a distance of 2000 bp. We find that $\sim 12\%$ of the top $1,484$ enhancers have an overlap with pro-

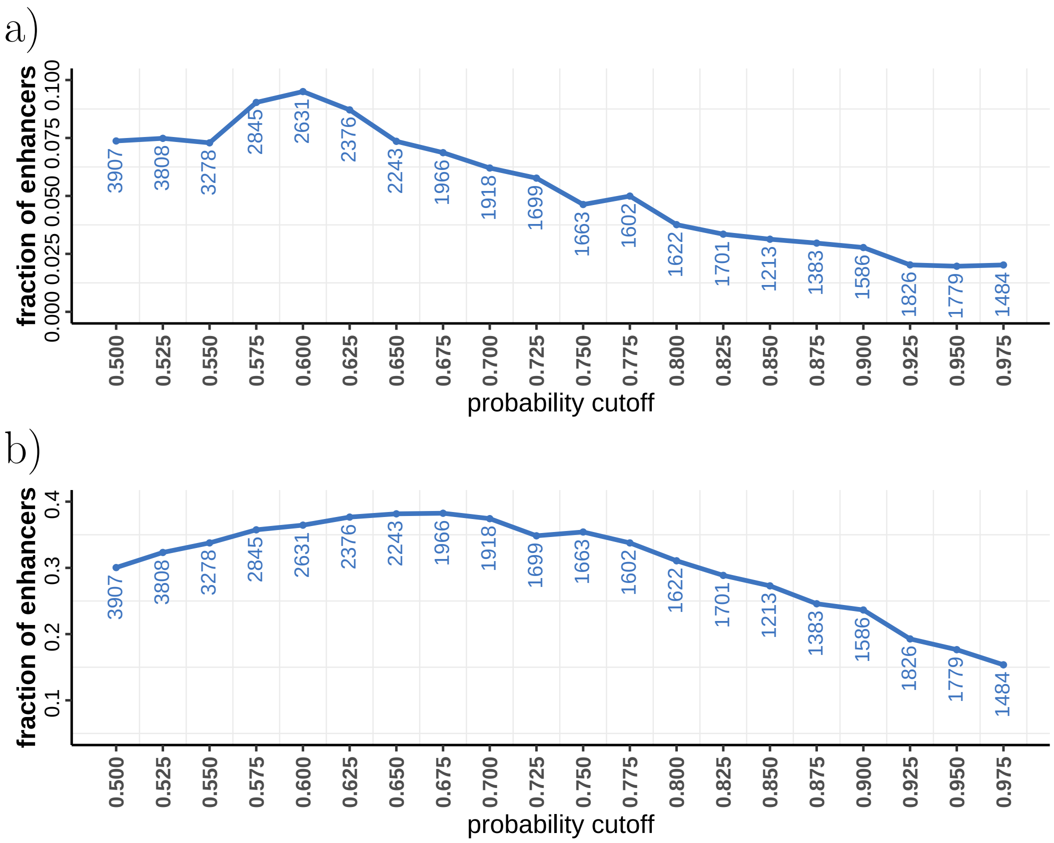


Figure 5.6: Percentage of promoter-proximal enhancers. We annotate enhancers in mESC with a probability threshold of 0.5 as described Section 4.1.7. For all enhancers falling into the same probability interval of size 0.025, for example the 1484 predicted enhancers in interval $[0.975, 1]$, we plot the fraction of enhancers that are close to an annotated promoter ($52,636$ TSSs in total). Promoter-proximal enhancers are defined based on a **a)** 200 bp or **b)** 2000 bp distance to the nearest promoter.

moters according to this criterion (see Figure 5.6 b). However, we also detect that there is more than one accessible region (in terms of ATAC-seq peaks) in the exact same vicinity for $\sim 80\%$ of the possible FPs. Hence, it is a likely scenario, that one of these accessible regions belongs to an active promoter, and another to the actual enhancer. The same could be observed for the second most confident set of enhancers, where $\sim 75\%$ of the overlapping enhancers have multiple ATAC-seq peaks nearby. The fraction of enhancers within a range of 2000 bp of a TSS stabilizes at around 33%. Taking into account the HM levels and the chromatin accessibility, we could divide the set of $42,530$ enhancers with a probability ≥ 0.5 into two clusters: a smaller cluster dominated by promoter-associated H3K4me3, and a big cluster where H3K4me1 is more enriched than H3K4me3 (see Figure 5.7 b). Especially for the smaller cluster, we observe multiple ATAC-seq peaks in the averaged profile supporting that only a fraction of enhancers in this cluster are FPs.

5.1.5 Prediction results for validated enhancer regions

To validate our classifier on a set of known enhancers independent of FANTOM5 data, we download a list of 25 mESC enhancers from Chen *et al.* (2008). These enhancers were originally identified based on simultaneous binding of the TFs *Nanog*, *Oct4* and *Sox2* and positively tested for activity using luciferase reporter assays transfected into ES cells. We make predictions on the central 100 bp bin of the validated enhancers and achieve very high probabilities > 0.72 for 23 of the 25 regions (see Figure 5.8 a)). Interestingly, all 25 validated enhancers overlap with our $42,530$ predicted enhancers which we called based on a 0.5 probability cutoff (see also Section 5.1.3). The corresponding probabilities are > 0.6 for all predicted enhancers, as can be seen in Figure 5.8 b).

5.1.6 Comparison of predicted enhancer clusters to annotated super enhancers

Based on the approach described in Section 4.1.8 we reduced our $42,530$ predicted enhancers to $7,550$ clusters or domains of high enhancer density, and $9,170$ single enhancers which do not have any enhancers in their vicinity (12.5 kb). Then, we compared our enhancer clusters to 927 annotated mESC super-enhancers (SEs) from Novo *et al.* (2018), which were defined mainly based on H3K27ac (see Section 2.4.9 for general information about SEs). We found that 96% (or 896) of the annotated SEs overlap with our enhancer clusters and all but three show an overlap with our complete list of predicted enhancers. Two

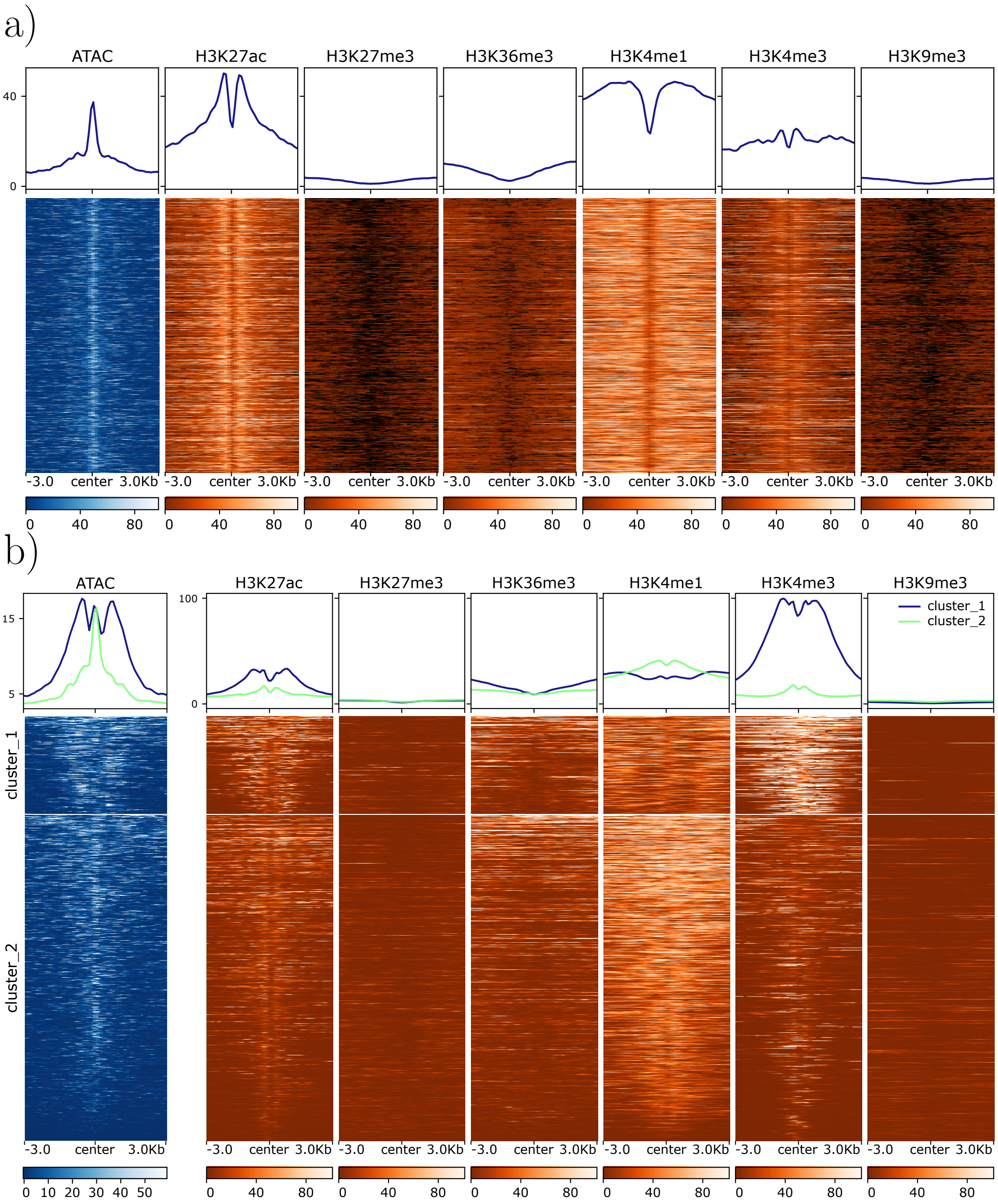


Figure 5.7: Histone modifications and accessibility at predicted enhancers. **a)** Top 1,484 predicted enhancers with probabilities > 0.975. **b)** All 42,530 predicted enhancers with default cutoff of 0.5 clustered with *deepTools* (Ramírez *et al.*, 2016). First cluster shows high promoter-associated mark H3K4me3 and ATAC-seq profile with multiple peaks. Second cluster with enhancer-associated HM patterns, e.g., higher levels of H3K4me1 and single ATAC-seq peak.

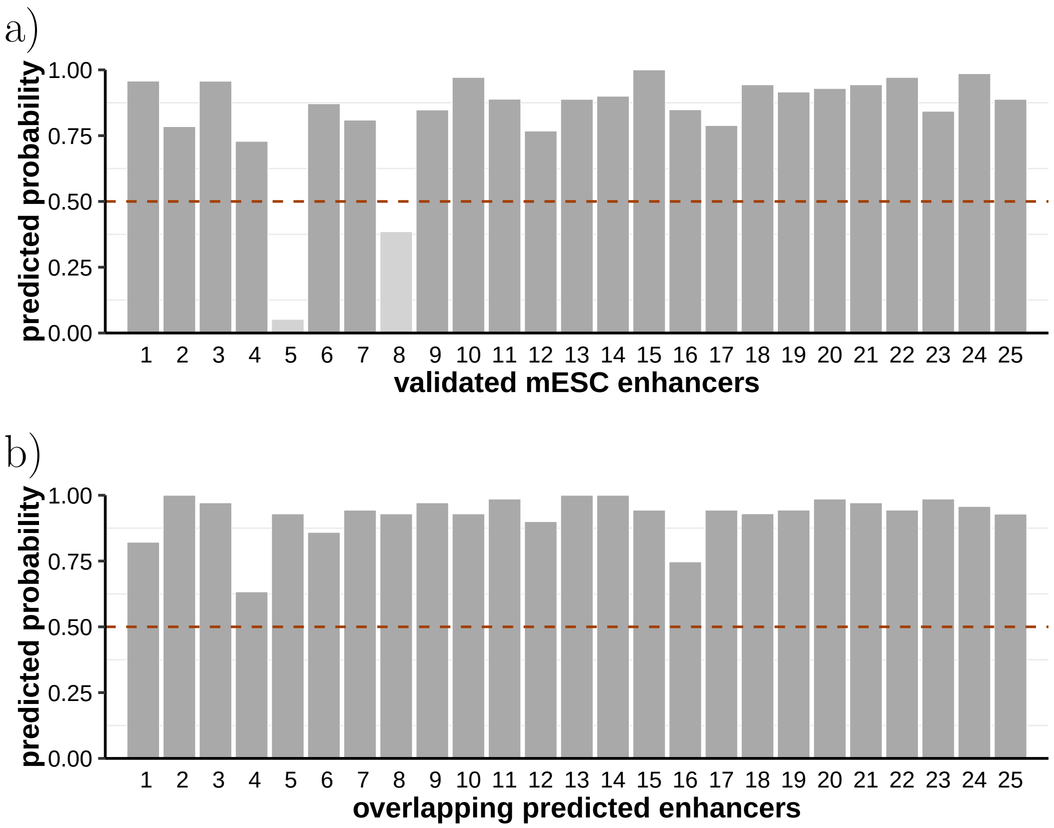


Figure 5.8: Predicted probabilities for validated enhancers. **a)** Predicted probabilities on the central 100 bp bin of the 25 validated mESC enhancers from Chen *et al.* (2008). Nearly all (23/25) enhancers have a predicted probability over our default cutoff of 0.5 (orange dashed line). **b)** Predicted probabilities of the 25 called enhancers that overlap with the set of validated enhancers.

example regions can be in Figure 5.9.

5.1.7 Motif analysis

We performed a motif analysis according to Section 4.4 on the total set of predicted enhancers (42, 530), as well as on the most confident top 1, 484 enhancers which have a probability > 0.98. Note, that we used the consensus motifs from the JASPAR database which were computed by a clustering approach on the whole set of individual motifs (Khan *et al.*, 2018).

To cover mostly the accessible part at the center of an enhancer and not the

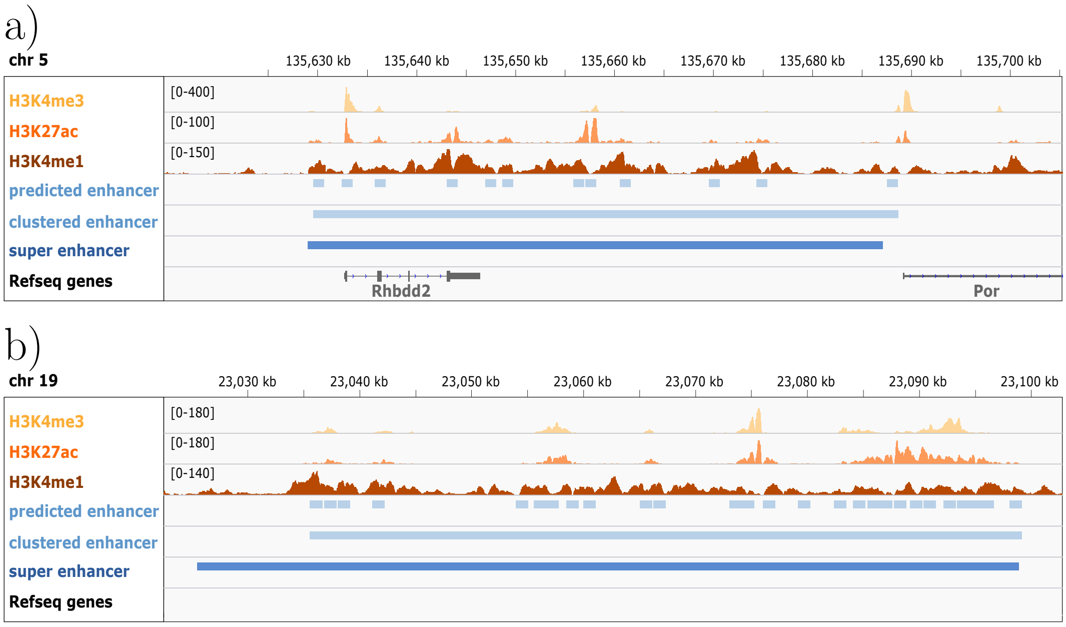


Figure 5.9: Overlap with super enhancers. IGV browser shot of two regions in the $mm10$ genome showing three HM profiles from the mESC data set, our predicted enhancer annotation ('predicted enhancer'), predicted domains of high enhancer density ('clustered enhancer'), annotated super enhancers from Novo *et al.* (2018) ('super enhancer') and annotated genes ('Refseq genes').

chromatin occupied by nucleosomes, we reduce each enhancer to 300 bp. This choice is based on previous results, for example regarding feature importance or hyperparameter optimization (see Sections 5.1.1 and sec:gridsearch). The background model for our motif analysis is computed from the total set of enhancers to take into account the underlying sequence composition of enhancers in contrast to random sequences (e.g., long stretches of repeats).

Since the set of all active enhancers in mESC is not specific but likely includes, for example, ubiquitously active enhancers, it is not straightforward to interpret the results of the motif analysis. The three most enriched TFs for both sets are *EWSR1-FLI1/ ZNF263*, *ZNF384* and *RREB1* (see Figure 5.10 which are, to our knowledge, not directly associated to mESCs. The motif logos are depicted in Figure 5.11. As could be expected for enhancers directly associated to the pluripotent state, the enhancer sets show high enrichment ($\geq$ 2 fold enrichment) for the known pluripotency factors *POU5F1*

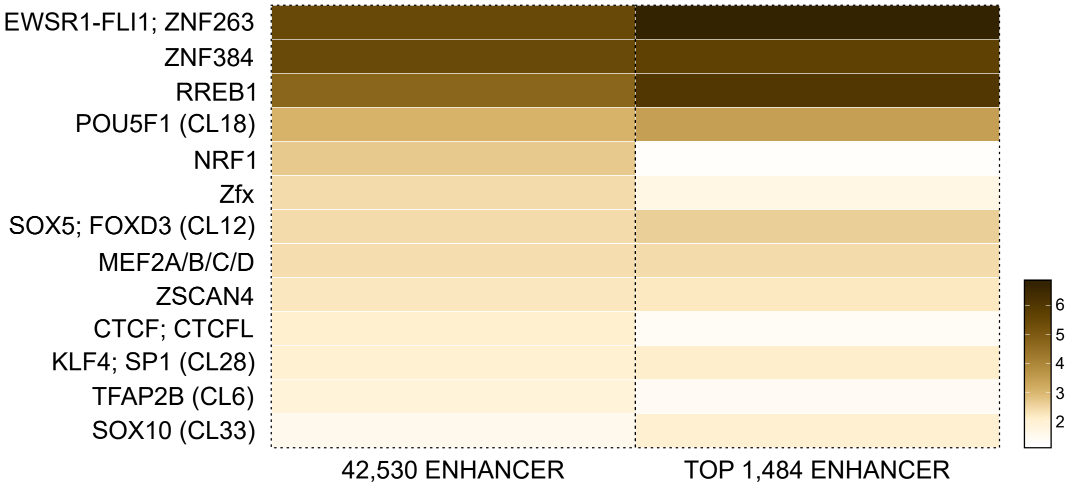


Figure 5.10: Motif enrichment in mouse embryonic stem cell enhancers. Motif enrichment values (fold-enrichment) in two sets of mouse embryonic stem cell enhancers. Depicted are only JASPAR consensus motifs with enrichment > 2 in at least one of the two sets. For each motif cluster ('CL') only a subset of contained motifs is shown. The complete cluster summary can be found in Table B4.

(*OCT4*) and *SOX*2 (Boyer *et al.*, 2005), and also for *KLF4*, which is thought to play an essential role in ESC self-renewal by regulating the gene expression of *Nanog* (Zhang *et al.*, 2010). The motif logos of *POU5F1* and also of the heterodimer *POU5F1::SOX2* are depicted in Figure 4.11. Interestingly, other highly enriched motifs in our enhancers sets, *MEF2*, *ZSCAN4* and *TFAP2B*, are associated to embryonic development (Lin *et al.*, 1998; Zalzman *et al.*, 2010; Moser *et al.*, 1997).

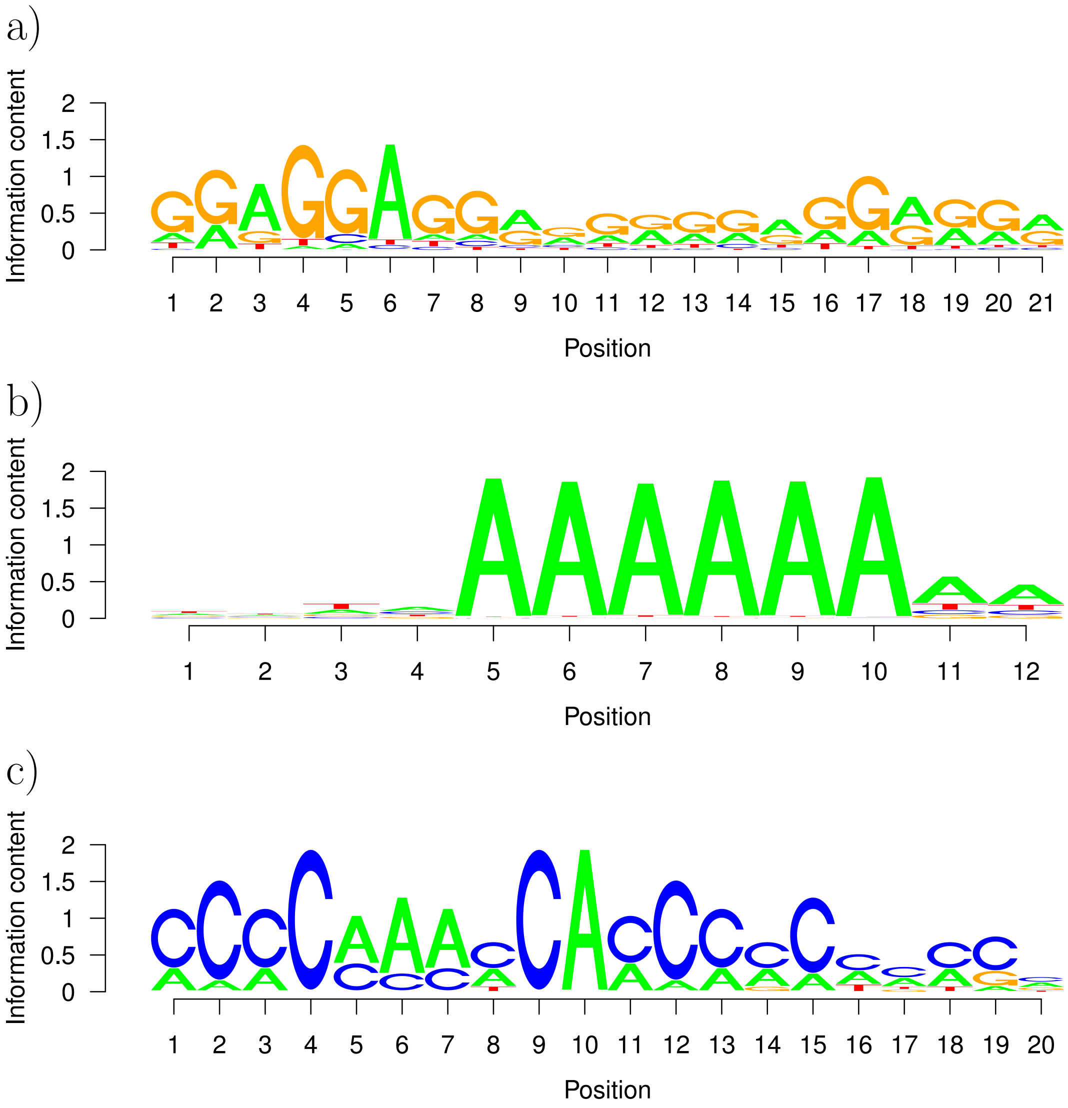


Figure 5.11: JASPAR motif logos for top 3 enriched motifs in mESC enhancers. **a)** consensus motif of cluster 54 **b)** consensus motif of cluster 55 **c)** consensus motif of cluster 75 Plot is done with R package *seqLogo* (Bembom, 2018) from the JASPAR pfms (normalized s.t. columns sum up to 1)

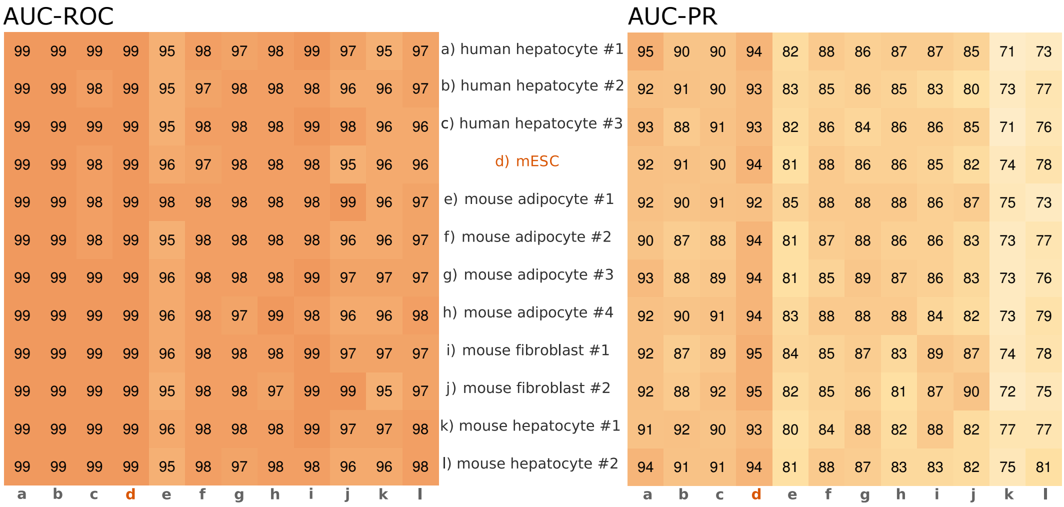
5.2 Enhancer predictions across cell types and species

So far, we were able to show that we can reliably predict enhancers genome-wide when training and predicting in the same cell type. A more difficult but also more realistic challenge is to predict enhancers in conditions or cell types for which there is no labeled training set available. Therefore, we analyze how well we can annotate enhancers across different cell types and even different species using a pre-trained enhancer classifier from a different setting.

For this purpose, we use in addition to mESC also three other cell types in mouse (fibroblasts, adipocytes, hepatocytes), as well as a human hepatocytes data set. Including replicates, this results in 12 different data sets which are summarized in Table 5.1. We integrate ChIP-seq, RNA-seq and DNase-seq data for all samples into the corresponding classification models. In the Appendix (Data Processing), we describe in more detail how the data was prepared and processed.

Table 5.1: Summary of data samples.

Abbreviation	Species	Tissue/Cell type
mESC	mouse	ESC blastocyst
mouse fibroblast #1 ('Mf05')	mouse	synovial fibroblast
mouse fibroblast #2 ('Mf07')	mouse	synovial fibroblast
mouse adipocyte #1	mouse	adipocyte/white fat cell
mouse adipocyte #2	mouse	adipocyte/white fat cell
mouse adipocyte #3	mouse	adipocyte/white fat cell
mouse adipocyte #4	mouse	adipocyte/white fat cell
mouse hepatocyte #1	mouse	liver hepatocyte
mouse hepatocyte #2	mouse	liver hepatocyte
human hepatocyte #1	human	liver hepatocyte
human hepatocyte #2	human	liver hepatocyte



AUC-ROC

	a	b	c	d	e	f	g	h	i	j	k	l
a) human hepatocyte #1	99	99	99	99	95	98	97	98	99	97	95	97
b) human hepatocyte #2	99	99	98	99	95	97	98	98	98	96	96	97
c) human hepatocyte #3	99	99	99	99	95	98	98	98	99	98	96	96
d) mESC	99	99	98	99	96	97	98	98	98	95	96	96
e) mouse adipocyte #1	99	99	98	99	98	98	98	98	98	99	96	97
f) mouse adipocyte #2	99	99	98	99	95	98	98	98	98	96	96	97
g) mouse adipocyte #3	99	99	99	99	96	98	98	98	99	97	97	97
h) mouse adipocyte #4	99	99	99	99	96	98	97	99	98	96	96	98
i) mouse fibroblast #1	99	99	99	99	96	98	98	98	99	97	97	97
j) mouse fibroblast #2	99	99	99	99	95	98	98	97	99	99	95	97
k) mouse hepatocyte #1	99	99	99	99	96	98	98	98	99	97	97	98
l) mouse hepatocyte #2	99	99	99	99	95	98	97	98	98	96	96	98

AUC-PR

	a	b	c	d	e	f	g	h	i	j	k	l
a) human hepatocyte #1	95	90	90	94	82	88	86	87	87	85	71	73
b) human hepatocyte #2	92	91	90	93	83	85	86	85	83	80	73	77
c) human hepatocyte #3	93	88	91	93	82	86	84	86	86	85	71	76
d) mESC	92	91	90	94	81	88	86	86	85	82	74	78
e) mouse adipocyte #1	92	90	91	92	85	88	88	88	86	87	75	73
f) mouse adipocyte #2	90	87	88	94	81	87	88	86	86	83	73	77
g) mouse adipocyte #3	93	88	89	94	81	85	89	87	86	83	73	76
h) mouse adipocyte #4	92	90	91	94	83	88	88	88	84	82	73	79
i) mouse fibroblast #1	92	87	89	95	84	85	87	83	89	87	74	78
j) mouse fibroblast #2	92	88	92	95	82	85	86	81	87	90	72	75
k) mouse hepatocyte #1	91	92	90	93	80	84	88	82	88	82	77	77
l) mouse hepatocyte #2	94	91	91	94	81	88	87	83	83	82	75	81

Figure 5.12: Test set performance across different cell types and species. Area under ROC curve (AUC-ROC) and area under precision recall curve (AUC-PR) results for $12 \cdot 12 = 144$ test set predictions. Classifiers are trained based on data indicated in the rows of the matrices, and tested on data indicated in the columns.

5.2.1 Area under ROC and precision-recall curve

For each sample, we first construct a training and an independent test set according to the descriptions in Section 4.1.3. Promoters and active enhancers are defined incorporating DNase-seq, RNA-seq and FANTOM5 data, where details can be found in Sections 4.1.5 and 4.1.6, as well as in Tables B1, B2 and B3. Then, we train 12 classifiers, one for each data set, using the same six core HMs as in the previous chapter (H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K36me3, H3K9me3) and make predictions on the test sets of all samples. Here, an important step is to normalize the (already input-normalized) HM count values of the test set according to the distribution of the training set as discussed further in Section 5.2.3. Finally, we compute the area under the ROC curve (AUC-ROC) and the area under the precision recall curve (AUC-PR) for all $12 \cdot 12 = 144$ test set predictions, depicted in Figure 5.12.

The AUC-ROC results are very high for all cell types, but less reliable, since the test sets are unbalanced with an estimated 10% true active enhancers. Therefore, we mostly concentrate on the AUC-PR as a measure of performance here. The AUC-PR over all training set and test set predictions are good with val-

ues $\in [0.71, 0.95]$. As a comparison, the expected result for a random classifier would be 0.1 according to the fraction of true enhancers in the test set. Interestingly, the performance seems to depend rather on the test than on the training set origin, with best results on the human hepatocyte and the mESC test sets. This could be due to differences in the HM data quality, e.g., sparse signals for the most important features (see also Figure A6 for data quality). Another reason could be the test set quality in terms of wrongly labeled samples. As already discussed in Section 5.1.2, both enhancers and non-enhancers can be mislabeled, for example due to not stringent enough cutoffs. While a few wrongly labeled enhancers (false positives) in the training set can be tolerated and the HM cutoffs can still be learned reliably, every false label in the test set immediately influences the reported performance. Following this logic, it makes sense that the best performances could be achieved on the test sets of mESC and human hepatocytes, since for both data sets the training enhancers were chosen more stringently (see Table B1).
It is unlikely, that the performance dependence on the test set is a sign of overfitting, which means learning to predict just a certain kind of enhancer. For example, if all training enhancers of a certain cell type are very active with appropriately high values of H3K27ac, the classifier cannot correctly learn to predict enhancers with lower activity level. However, if this would be the case here, we we would expect to observe high results in the diagonal which represents the prediction performance when training and test set data are from the same sample origin and hence chosen according to the same criteria. As can be seen in Figure 5.12, the diagonal entries rarely give the best result. Hence, we suspect that the classifiers are able to learn how to predict enhancers from all individual training sets, but certain regions in the test set cannot be labeled correctly independently of classifier transferability from one cell type or species to another.

5.2.2 Pre-trained optimized enhancer classifier

For further enhancer predictions in cell types without available training set, we prepared a pre-trained classifier. Based on the high quality of the training set and taking into account the performance results of the previous section, we chose to use the pre-trained mESC classifier for this purpose. Details about how the classifier was trained and evaluated within mESC can be found in Section 5.1. In summary, the classifier we are offering is based on ChIP-seq data from six core HMs (H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K36me3 and H3K9me3) as well as a control (Input) sample, and was trained on 1000 genomic regions from which 100 are active enhancers. The used enhancers are shown to produce bidirectional transcripts in mESCs and are centered at their accessible region.

To make predictions, the user only has to provide the ChIP-seq data. We input-normalize each of the six HM samples and then use quantile normalization with the original mESC data as reference. Both normalization steps (input and quantile normalization) are integrated into our enhancer prediction method and thus do not have to be carried out by the user. We discuss the necessity of the quantile normalization step in more detail in the following section.

5.2.3 Quantile normalization

Before applying our pre-trained classifier on a new data set, we make two types of normalization. First, we input-normalize each HM sample as described in Section 4.1.1 to account for a non-uniform background read distribution. Then, we apply quantile normalization on these values taking the training data, i.e., the input-normalized HMs from our mESC sample, as a reference (*normalize.quantiles.target* function from the R package *preprocessCore*, Bolstad (2018)). This equalizes the HM distributions in terms of their statistical properties. Since we use cutoff-based random forests in our classification model, it is important to have comparable HM values between our training sample and the sample we want to predict on, which could be a replicate or equally well from a different cell type.

To check the effects of quantile normalization on samples with varying coverage, i.e., varying numbers of unique sequenced reads, we applied our pre-trained classifier to our mESC data set where we randomly subsampled proportions $(10\%, 20\%, \ldots 90\%)$ of the alignments using *SAMtools* on each HM and the input data (Li *et al.*, 2009). If we do not quantile-normalize the subsampled data sets with respect to the original training data, we get a higher frequency of genomic bins with predicted enhancer probabilities $\in [0.15, 0.65]$, and fewer bins with high probabilities (see Figure 5.13 a)), which leads to decreasing numbers of genome-wide predicted enhancers (e.g., for the default probability cutoff 0.5). This trend is more pronounced as the proportion of subsampled reads decreases. Thus, differences in coverage seem to be directly reflected in the distribution of predicted probabilities which makes it more difficult to compare results from different samples regarding the number of predicted enhancers as well as the bin-wise enhancer probabilities. If however we include the step of quantile normalization into our workflow, we can adjust the enhancer probabilities towards the predictions on the original data set (see Figure 5.13 b)). We lose the elevated frequencies of probabilities $\in [0.15, 0.65]$ that we saw without quantile normalization, and also the frequencies of high probability bins is much higher and more comparable.

To see the effects across cell types, we applied our pre-trained mESC classifier with and without quantile normalization on two biological replicates of mouse synovial fibroblast data ('Mf05' and 'Mf07' in Figure 5.14). Interestingly, the two biological replicates seem to be quite different regarding the probability distribution independent of normalization, which is likely due to quality differences of the HM ChIP-seq data. In Figure 5.15, we plot the cumulative sum of reads per 500 bp bin according to their ranks and see strong differences for the H3K27ac curves of the fibroblast samples. The results for 'Mf05' look as expected, with a steep slope towards the highest ranks indicating specific enrichment across the genome, but the curve for 'Mf07' is close to the diagonal which means that the reads are nearly uniformly distributed. While these differences in data quality seem to not be resolvable with quantile normalization, we observe that the frequency distribution of predicted probabilities $\in [0.05, 0.35]$ are comparable between both fibroblast and the mESC data only if we include the step of quantile normalization. Furthermore, the probability distribution of the fibroblast data with better quality is overlapping with the mESC distribution even until probabilitites of ~ 0.75, and afterwards it is close.

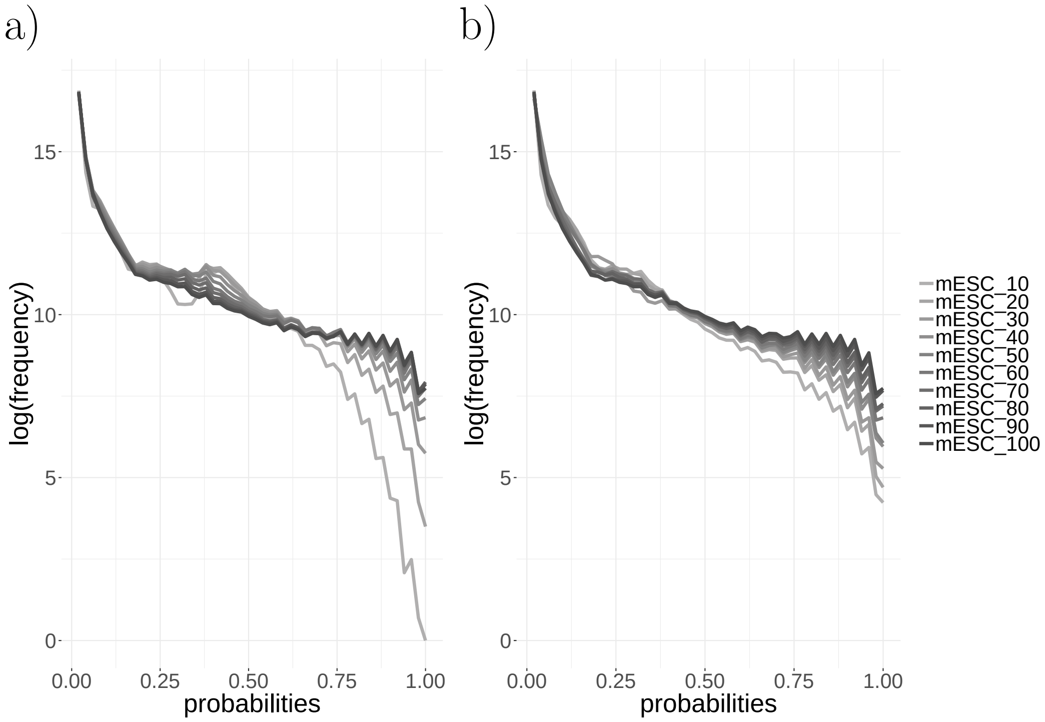


Figure 5.13: Effects of quantile normalization for varying coverage. We applied the pre-trained mESC classifier from Section 5.2.2 to the mESC data set with varying coverage **a)** without quantile normalization, or **b)** with quantile normalization, and predict enhancer probabilities genome-wide (in 100 bp bins). We plot the frequency of bins (log-scale) according to their predicted enhancer probability. 'mESC_10' corresponds to prediction results based on an mESC data set where we subsampled 10% of the original alignment.

5.3 Comparison to other enhancer prediction methods

Genome-wide identification of enhancers has been a topic of interest for decades and led to many published enhancer prediction methods based on different computational approaches and input data types. However, the existing methods still have limitations which can be due to data availability or method usability, especially when wanting to make predictions on new data sets (see Kleftogiannis *et al.* (2016) for an overview of methods and challenges). The majority of existing methods are either based on unsupervised HMM-based al-

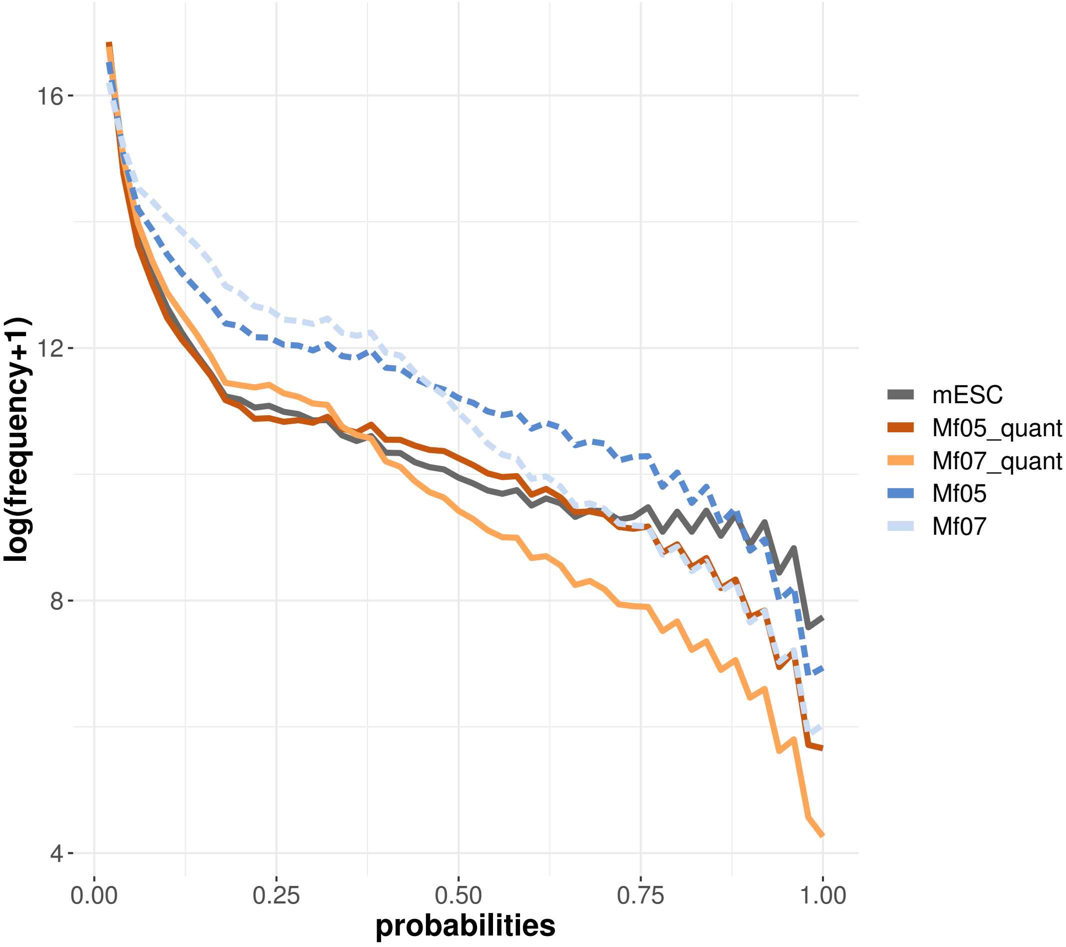


Figure 5.14: Effects of quantile normalization across cell types. We used the pre-trained mESC classifier from Section 5.2.2 to predict enhancer probabilities genome-wide (in 100 bp bins) in the mESC sample and two mouse fibroblast samples (biological replicates). We made predictions using quantile normalization ('Mf05_quant' and 'Mf07_quant'), and without quantile normalization ('Mf05' and 'Mf07').

gorithms or supervised approaches. In this work, we compare our enhancer prediction method to the unsupervised genome segmentation tool ChromHMM, which is extensively used for enhancer prediction due to its flexibility and easy usability (Ernst and Kellis, 2012, 2017). Furthermore, we make a comparison to the recently published supervised method REPTILE which reached a superior performance over many other state-of-the art methods for enhancer

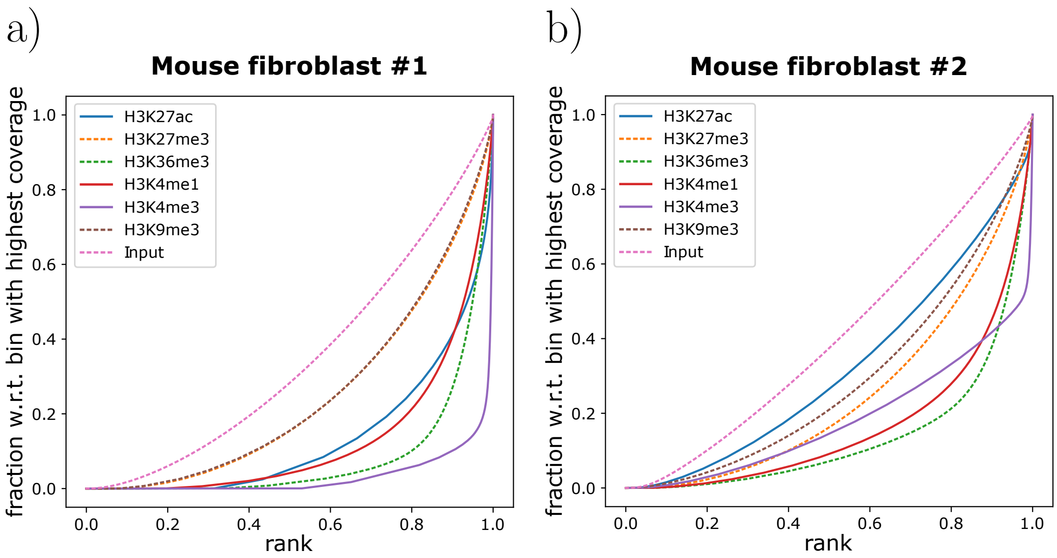


Figure 5.15: Fingerprint quality control metrics for mouse fibroblast ChIP-seq experiments. For each HM ChIP-seq data, reads with a mapping quality ≥ 30 are counted per adjacent 500 bp bin. Then, the read counts are sorted and their cumulative sum is plotted. The plots are done with *deepTools* (Ramírez *et al.*, 2016).

prediction (He *et al.*, 2017). Due to the different objective of ChromHMM (genome segmentation), we solely compare it to our method based on its performance on a test set. REPTILE, however, has the same objective and output as our enhancer prediction method (predicted probabilities and annotated enhancer candidates) and can therefore be evaluated based on the same standards: performance on a test set, spatial resolution of the annotated enhancers, overlap with promoters, as well as transferability to new data sets. A detailed summary of both approaches can be found in Section 4.5.

ChromHMM

As described in Section 4.5.1, ChromHMM is a genome-segmentation tool which is based on hidden Markov models (Ernst and Kellis, 2012, 2017). We applied ChromHMM to our six core HMs and three different state parameters, $K = 8, 12, 16$. The genome wide segmentation into 8 states did not result in a clear enhancer state according to the emission probabilities. As can be

seen in Figure 5.16 for $K = 8$, the only state with high probabilities for the enhancer-associated marks H3K27ac and H3K4me1 (state $E2$), also shows the promoter-associated H3K4me3. For further analysis, we therefore concentrate on chromatin segmentations into 12 and 16 states. For $K = 12$, the state $E3$ shows clear enhancer marks, since only H3K27ac and H3K4me1 have high emission probabilities. We also expect $E2$ to cover several active enhancers as well as active promoters due to a high H3K4me1 and H3K4me3 probabilities. $E12$ could contain weak enhancers in intragenic regions indicated by the transcriptional elongation mark H3K36me3. The genome wide segmentation into $K = 16$ chromatin states results in a similarly clear active enhancer state ($E1$) as well as a mixed enhancer-promoter state ($E6$) and a state that could cover weak intragenic enhancers ($E16$).
Based on these observations, we define for $K = 12$ either

- $E3$ as 'enhancer' and all other states as 'non-enhancer',
- $E3 + E12$ as 'enhancer' and all other states as 'non-enhancer',
- $E3 + E2$ as 'enhancer' and all other states as 'non-enhancer',
- or $E3 + E2 + E12$ as 'enhancer' and all other states as 'non-enhancer';

and for $K = 16$, following the same logic, either

- $E1$ as 'enhancer' and all other states as 'non-enhancer',
- $E1 + E16$ as 'enhancer' and all other states as 'non-enhancer',
- $E1 + E6$ as 'enhancer' and all other states as 'non-enhancer',
- or $E1 + E6 + E16$ as 'enhancer' and all other states as 'non-enhancer'.

Then we make predictions on our ten test sets, which were described in Section 4.1.3 and measured the false positive rate (FPR), the true positive rate (TPR) as well as the precision. Here, each test set region (size 1100 bp) that has any overlap with an annotated ChromHMM enhancer is consequently predicted as an 'enhancer', and in case of no overlap as a 'non-enhancer'. The predictions on our test sets lead to strongly varying true positive rates $\in [0.2, 0.925]$, which interestingly separate into two clusters (see Figure 5.17). For both segmentations the incorporation of the mixed promoter-enhancer-state ($E2$

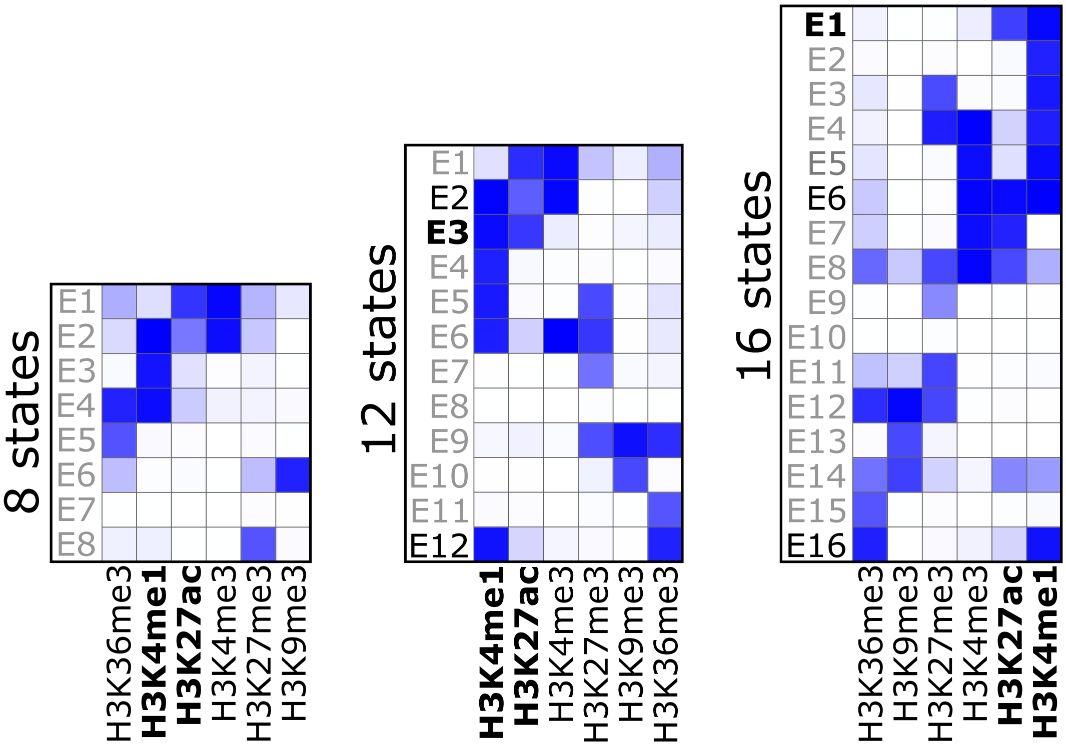


Figure 5.16: ChromHMM emission probabilities ChromHMM was applied on mESC data to make genome-wide segmentations into 8, 12 or 16 states. The heatmaps show emission probabilities in each state (rows) and each epigenetic mark (columns), where a dark colour corresponds to higher probabilities. H3K27ac and H3K4me1 (bold) are enhancer-associated marks, which are used to define exclusive enhancer states (highlighted in black and bold) and mixed enhancer states (highlighted in black).

for $K = 12$ and $E6$ for $K = 16$) leads to a huge increase in TPR performance. And since the TPR is the ratio between the number of all true positive (TP) predictions and the number of all actual positive regions in the test set, the two clusters indicate that many of the actual positive enhancers overlap with the mixed promoter-enhancer-state and were consequently missed when not included.

For the varying precision values $\in [0.36, 0.775]$ the inclusion of the promoter-like state increases the performance a little, while adding the intragenic (weak) enhancers has a decreasing effect on the performance. Here, we suspect that we add more false positive (FP) predictions than TPs ,since the precision is

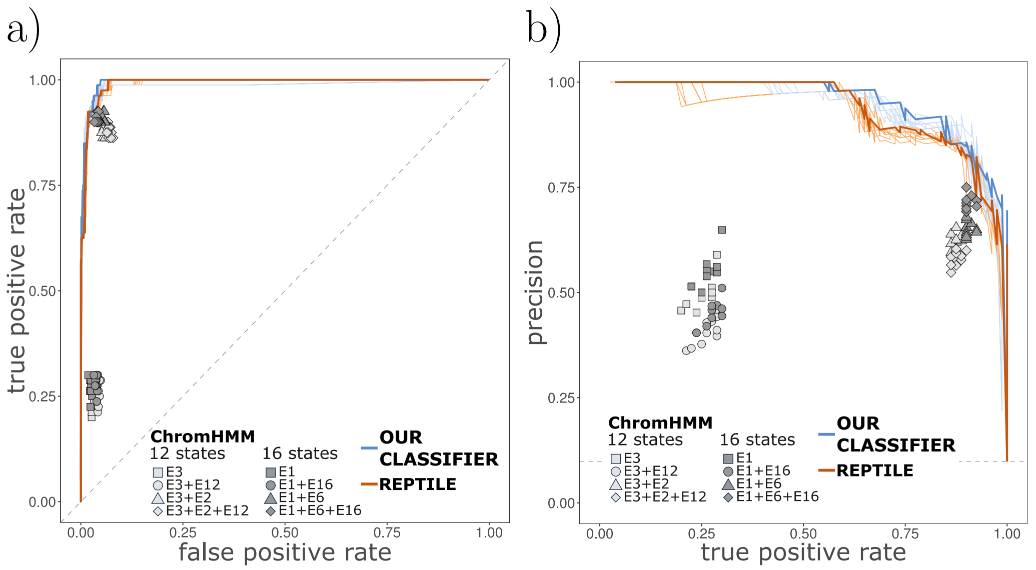


Figure 5.17: Comparison of ROC and precision-recall curve. We apply our classifier, ChromHMM and REPTILE to the mouse ESC data and make predictions on 10 (overlapping) test sets. ChromHMM was applied for $K = 8$ and $K = 12$ states, and for different definitions of the final enhancer state. REPTILE was trained with four different settings. Shown are **a)** ROC curves and **b)** precision-recall curves for our classifier and REPTILE. The curves with the highest area under the curve (AUC) are highlighted in darker colours. Since the output of ChromHMM is binary (and not a probability value), results for the 10 test sets are depicted as single instances and not as curves.

calculated based on only these two metrics (see also Section 3.3.4 for more details on the performance metrics). The FPR shows stable good results between 0.018 and 0.079.

Overall, these findings show that the performance results of ChromHMM greatly vary and therefore depend on the definition of the enhancer state which has to be done manually by the user (e.g. by eyeballing the emission probabilities). Additionally, there is no single state that uniquely describes enhancers. This emphasizes the difficulties in separating enhancers from active promoters. Our classifier clearly outperforms ChromHMM independent of its settings or the choice of enhancer states. For our default probability cutoff of 0.5, for example, our results of TPR $\in [0.85, 0.913]$, precision $\in [0.833, 0.869]$ and also FPR $\in [0.015, 0.019]$ (see Figure 5.17) are superior to ChromHMM for the purpose of enhancer prediction.

REPTILE

REPTILE, described in more detail in Section 4.5.2, is a supervised method designed specifically for enhancer prediction (He *et al.*, 2017). It is also based on random forests and, in its original setting, needs HM ChIP-seq and also methylation data as input. The software to train a classifier is freely availbale, as well as three pre-trained classifiers which were trained on p300 binding sites in mESC and

(i) the three HMs H3K27ac, H3K4me1 and H3K4me3,

(ii) the six HMs H3K27ac, H3K4me1, H3K4me3, H3K9ac, H3K36me3 and H3K27me3, or

(iii) six HMs and DNA methylation data.

The mESC ChIP-seq data used by He *et al.* (2017) is different from our data set, and will be called $\mathrm{mESC}_{\mathrm{REP}}$ from now on.

Performance on a test set

To measure the performance on test sets in the most comparable way, we trained REPTILE on the same data as our method, i.e., the FANTOM5-based training set described in Section 4.1.3 and the six core HMs in mESC. REPTILE could achieve high results in terms of area under the precision-recall curve, AUC-PR $\in [0.92, 0.94]$, as well as an AUC-ROC ~ 0.99 over all test sets (see Figure 5.17).

For a probability cutoff of 0.5, REPTILE achieves an FPR $\in [0.013, 0.018]$ and precision values $\in [0.843, 0.88]$ and a TPR $\in [0.825, 0.875]$. Based on the test set comparison, our method and REPTILE have a similar performance while our classifier is slightly better in terms of AUC-PR, and slightly worse looking at the ROC curve results. However, the AUC-PR is the more reliable quality measure since it is best suited for imbalanced test sets (see Section 3.3.4 for more details).

Spatial resolution

We measured the spatial resolution of REPTILE by computing the distance to the closest accessible region following the same logic as in Section 5.1.3. To make a fair comparison and understand more about the differences between our approach and REPTILE, we did this for four settings based on different training and feature sets:

(i) FANTOM5 derived enhancers, our six core mESC HMs,

(ii) p300 defined enhancers, our six core mESC HMs and intensity deviation,

(iii) p300 defined enhancers, six mESC_{REP} HMs, and

(iv) p300 defined enhancers, six mESC_{REP} HMs, DNA methylation and differentially methylated regions (DMRs).

The p300 defined enhancers, mESC_{REP} HM ChIP-seq data, the DNA methylation data, te DMRs and the intensity deviation features are used in the original REPTILE publication by He *et al.* (2017) and are also described in more detail in Section 4.5.2.

Taking a probability cutoff of 0.5, we called (i) 24,823, (ii) 34,584, (iii) 32,797 and (iv) 30,360 enhancer regions using the REPTILE pipeline described in Section 4.5.2, and $42,530$ enhancers with our classifier. Increasing the probability cutoff in steps of 0.025 until reaching a probability of 1, the number of predicted enhancers decreases until we only call the (i) 2,248, (ii) 1,814, (iii) 3,087 and (iv) 2,447 most confident enhancers with the REPTILE method and $1,484$ with ours. Then, we compute the distance to the closest ATAC-seq peak in mESC for each enhancer and compute the median of the distances for each enhancer set.

As a general trend, the median distance to the closest accessible region increases with decreasing confidence in the enhancer prediction, i.e., decreasing cutoff probability (see Figure 5.18). For the $2,000$ most reliably predicted enhancers, the difference in spatial resolution between our method (~ 120 bp) and the best REPTILE setting (~ 200 bp) is very high. Up until the top $12,000$ annotated enhancers, our classifier outperforms the other methods, with a median distance of ~ 250 bp. Until we reach $\sim 18,000$ predicted

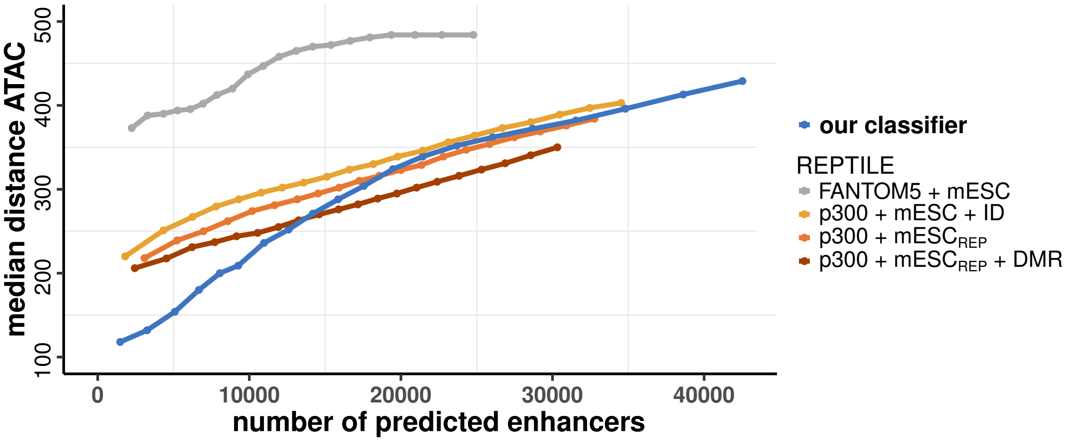


Figure 5.18: Distance to closest accessible region. We made genome-wide enhancer predictions on the mouse ESC data with our classifier and REPTILE for four different settings. Based on decreasing probability cutoffs from 1 to 0.5, we annotated different sized sets of enhancers. For each enhancer set, we computed the median of the distances between each annotated enhancer and its closest ATAC-seq peak.

enhancers, only additional use of DNA methylation data and DMRs leads to a better REPTILE performance. After that, our method is comparable to the settings (ii) and (iii) of REPTILE, which are based on the p300 training set. Enhancer predictions stemming from the REPTILE classifier trained on the FANTOM5-based enhancers and our mESC HM feature set, however, are on average 100 bp further away from an accessible region. To explore possible reasons for this observation, we measured the spatial resolution of our FANTOM5-based enhancers and of the p300-based enhancers from He *et al.* (2017). In Figure 5.19 a) it can be seen that the results are very similar, and that the majority of enhancers in both sets have a median distance from the closest ATAC-seq peak $\in [10, 100]$ bp.

Another possibility could lie in a bigger diversity of FANTOM5 versus p300-based enhancers in terms of their HM profile. Interestingly, our method has a clearly higher spatial resolution using the exact same training set and feature data (compare grey line in Figure 5.18). This shows that taking several neighboring bins into account, as well as combining two classifiers to decouple the distinction between enhancers and promoters, lead to a more precise location

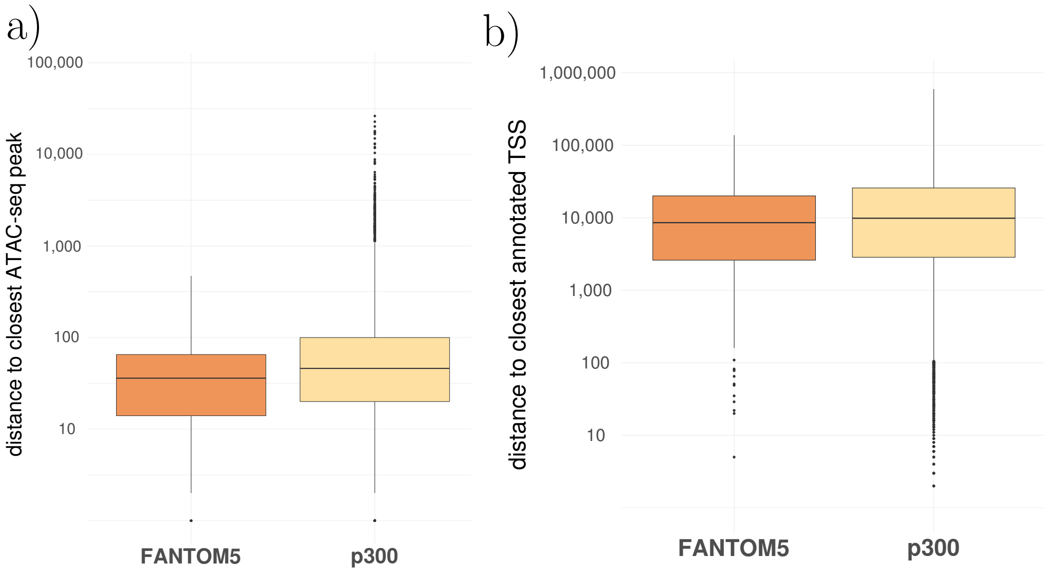


Figure 5.19: Distance to accessible region and transcription start sites for p300 and FANTOM5-based training enhancers. **a)** Boxplots summarize the the distances between enhancers and their closest ATAC-seq peak for FANTOM5-based enhancers (orange) and p300-based enhancers (yellow). **b)** Boxplots summarize the the distances between enhancers and their closest annotated transcription start site (TSS) for FANTOM5-based enhancers (orange) and p300-based enhancers (yellow). We define TSSs based on the Ensembl database (GRCh37.70).

of the predicted enhancers.

Promoter-proximal enhancer prediction

Here we measure the proximity or the overlap of predicted enhancers to annotated transcription start sites (TSSs) for the same four settings of REPTILE as described above. Depending on the distance, overlaps can indicate wrongly predicted enhancers that are actual active promoters.

First, enhancers are divided in subsets based on their probability, where we start with predicted probability values $\in [0.975, 1]$ and continue in steps of 0.05 until we reach probabilities $\in [0.975, 1]$ (see also to Section 5.1.4). This results in 20 different sets of enhancers with increasing confidence or predicted probabilities. Then, for each of these sets we compute the median of the fraction of enhancers that are not more than (i) 200 bp or (ii) 2000 bp away from the closest annotated TSS.

For the 200 bp distance, the REPTILE classifiers trained on the p300-based training enhancers have very stable and similar results, as can be seen in Figure 5.20 a). The classifier taking into account DMRs and adding DNA methylation as a feature shows the lowest overlap with promoters ($\sim 1.5\%$). In comparison to our method and the setting depending on FATOM5-based training enhancers, the fraction of likely false positives (FPs) is better. A potential reason could be a difference of the training enhancers attributed to the different properties used for their definition (binding of p300 vs. bidirection transcription).

The same analysis for a distance of 2000 bp, depicted in Figure 5.20 b), shows similar trends. Also here, the REPTILE classifiers trained on p300-based enhancers show a smaller overlap with promoters ($10-15\%$ for high-probability enhancers, and $15-25\%$ for enhancers with lower probabilities). Our classifier shows an overlap between 15% and 35%, and the REPTILE classifier trained on FANTOM5-based enhancers the highest with $35-50\%$.

As explained in more detail in Section 5.1.4, it is not certain that enhancers within a distance of 2000 bp of a TSS are really FP predictions or actual promoter-proximal enhancers. We investigate this possibility by computing the number of accessible regions within the same distance. If multiple accessible regions can be detected, this could hint towards the existence of both an active promoter and an actual active enhancer. Interestingly, we find that the fraction of enhancers close to a promoter and close to two or more ATAC-seq peaks is often higher for our method than for the best performing DMR-based REPTILE method (e.g. $\sim 78\%$ and $\sim 75\%$, respectively, for the most confident enhancers and $\sim 59\%$ and $\sim 55\%$ for the least confident ones). This may suggest, that we are able to predict more promoter-proximal enhancers than the competitor method.

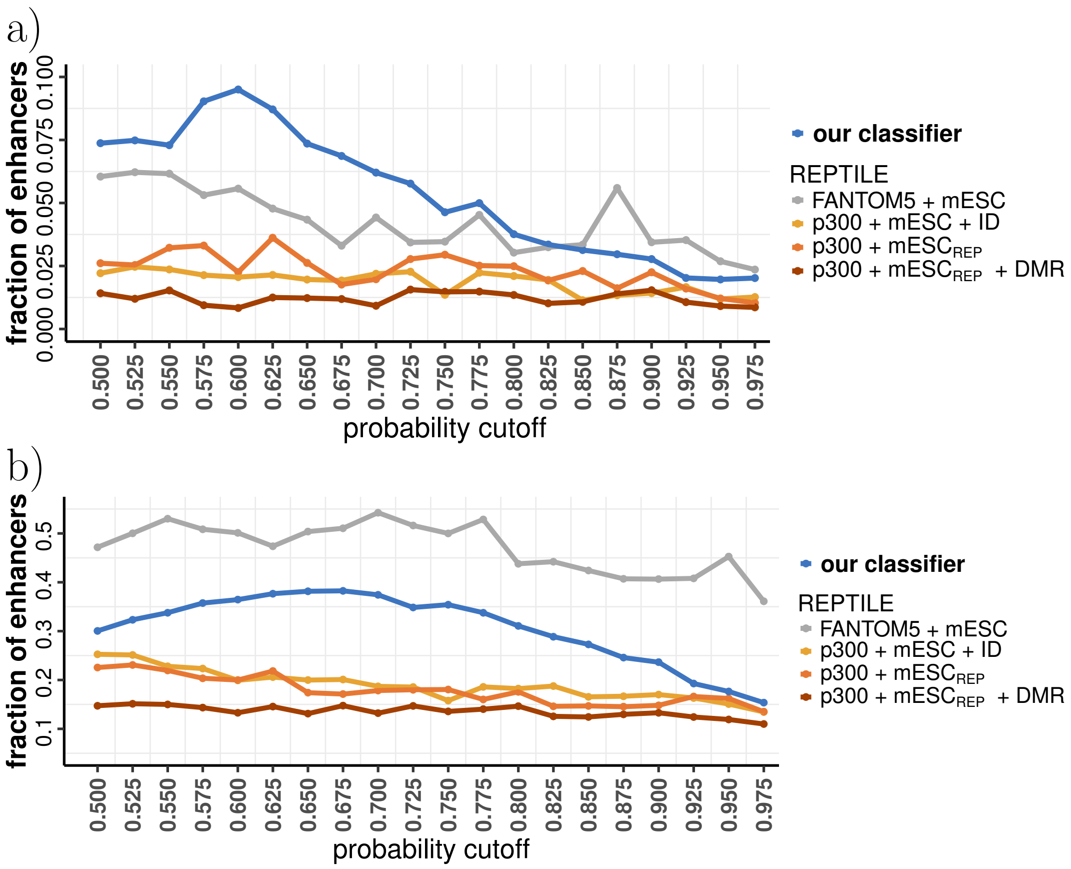
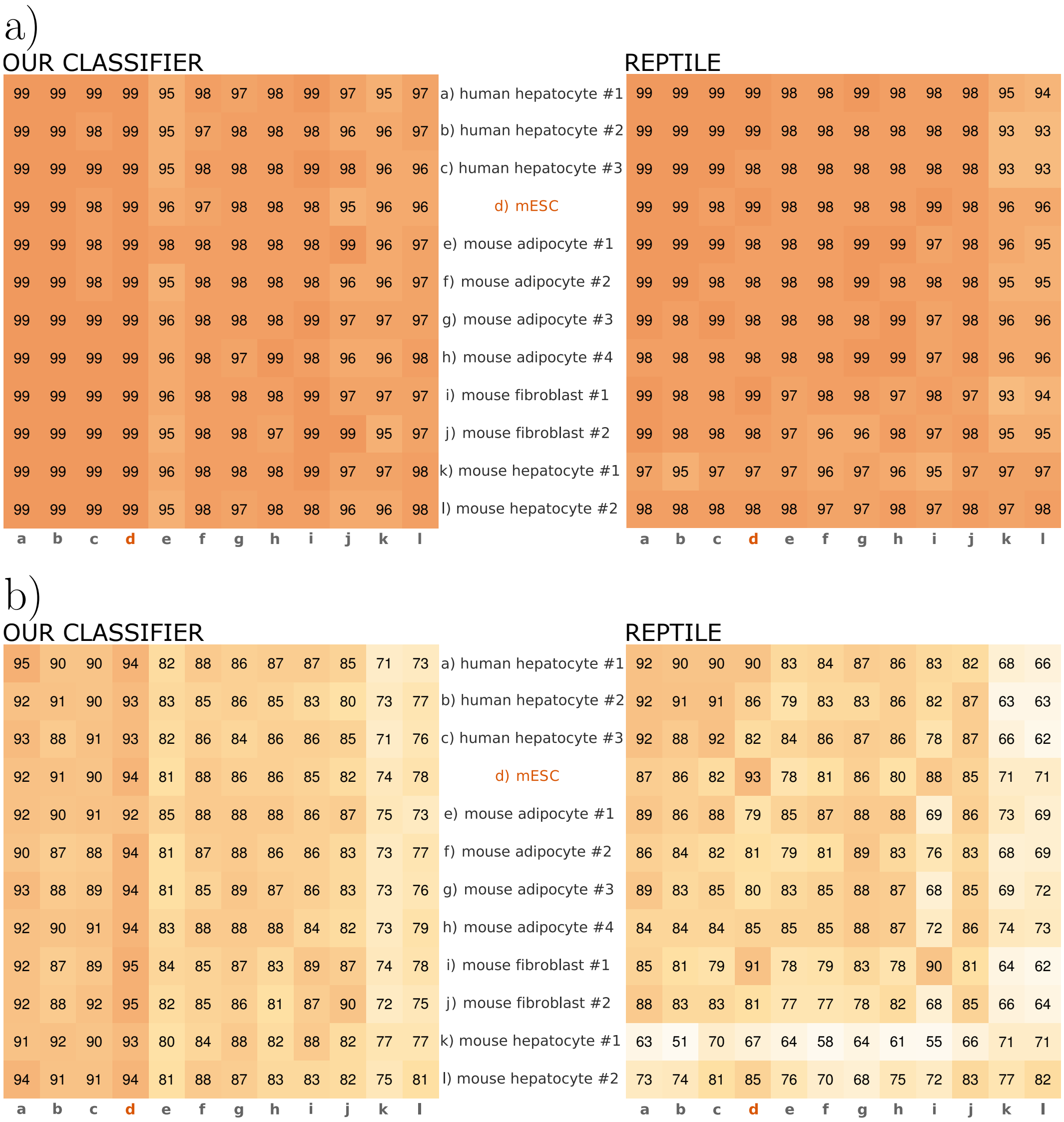


Figure 5.20: Percentage of promoter-proximal enhancers. We annotate enhancers in mESC with our method and REPTILE in four different settings with a probability threshold of 0.5. For all enhancers falling into the same probability interval, we plot the fraction of enhancers that are close to an annotated promoter (52,636 TSSs in total). Promoter-proximal enhancers are defined based on a **a)** 200 bp or **b)** 2000 bp distance to the nearest promoter.

Prediction across cell types and tissues

To see how well REPTILE can be applied across different cell types and species, we trained 12 classifiers on 12 different cell types in mouse and human using the FANTOM5-based training sets described in Section 4.1.3 and the HM data from Section 5.2. Then, we apply each of the classifiers on the remaining 11 data sets, and on an independent test set within the data set of training origin, to compute the AUC-ROC and AUC-PR. This results in 12 × 12 AUC-ROC and AUC-PR matrices depicted in Figure 5.21 a) and b), respectively.



a)

OUR CLASSIFIER

	a	b	c	d	e	f	g	h	i	j	k	l
a) human hepatocyte #1	99	99	99	99	95	98	97	98	99	97	95	97
b) human hepatocyte #2	99	99	98	99	95	97	98	98	98	96	96	97
c) human hepatocyte #3	99	99	99	99	95	98	98	98	99	98	96	96
d) mESC	99	99	98	99	96	97	98	98	98	95	96	96
e) mouse adipocyte #1	99	99	98	99	98	98	98	98	98	99	96	97
f) mouse adipocyte #2	99	99	98	99	95	98	98	98	98	96	96	97
g) mouse adipocyte #3	99	99	99	99	96	98	98	98	99	97	97	97
h) mouse adipocyte #4	99	99	99	99	96	98	97	99	98	96	96	98
i) mouse fibroblast #1	99	99	99	99	96	98	98	98	99	97	97	97
j) mouse fibroblast #2	99	99	99	99	95	98	98	97	99	99	95	97
k) mouse hepatocyte #1	99	99	99	99	96	98	98	98	99	97	97	98
l) mouse hepatocyte #2	99	99	99	99	95	98	97	98	98	96	96	98

REPTILE

	a	b	c	d	e	f	g	h	i	j	k	l
a) human hepatocyte #1	99	99	99	99	98	98	99	98	98	98	95	94
b) human hepatocyte #2	99	99	99	99	98	98	98	98	98	98	93	93
c) human hepatocyte #3	99	99	99	98	98	98	98	98	98	98	93	93
d) mESC	99	99	98	99	98	98	98	98	99	98	96	96
e) mouse adipocyte #1	99	99	99	98	98	98	99	99	97	98	96	95
f) mouse adipocyte #2	99	99	98	98	98	98	99	98	98	98	95	95
g) mouse adipocyte #3	99	98	99	98	98	98	98	99	97	98	96	96
h) mouse adipocyte #4	98	98	98	98	98	98	99	99	97	98	96	96
i) mouse fibroblast #1	99	98	98	99	97	98	98	97	98	97	93	94
j) mouse fibroblast #2	99	98	98	98	97	96	96	98	97	98	95	95
k) mouse hepatocyte #1	97	95	97	97	97	96	97	96	95	97	97	97
l) mouse hepatocyte #2	98	98	98	98	98	97	97	98	97	98	97	98

b)

OUR CLASSIFIER

	a	b	c	d	e	f	g	h	i	j	k	l
a) human hepatocyte #1	95	90	90	94	82	88	86	87	87	85	71	73
b) human hepatocyte #2	92	91	90	93	83	85	86	85	83	80	73	77
c) human hepatocyte #3	93	88	91	93	82	86	84	86	86	85	71	76
d) mESC	92	91	90	94	81	88	86	86	85	82	74	78
e) mouse adipocyte #1	92	90	91	92	85	88	88	88	86	87	75	73
f) mouse adipocyte #2	90	87	88	94	81	87	88	86	86	83	73	77
g) mouse adipocyte #3	93	88	89	94	81	85	89	87	86	83	73	76
h) mouse adipocyte #4	92	90	91	94	83	88	88	88	84	82	73	79
i) mouse fibroblast #1	92	87	89	95	84	85	87	83	89	87	74	78
j) mouse fibroblast #2	92	88	92	95	82	85	86	81	87	90	72	75
k) mouse hepatocyte #1	91	92	90	93	80	84	88	82	88	82	77	77
l) mouse hepatocyte #2	94	91	91	94	81	88	87	83	83	82	75	81

REPTILE

	a	b	c	d	e	f	g	h	i	j	k	l
a) human hepatocyte #1	92	90	90	90	83	84	87	86	83	82	68	66
b) human hepatocyte #2	92	91	91	86	79	83	83	86	82	87	63	63
c) human hepatocyte #3	92	88	92	82	84	86	87	86	78	87	66	62
d) mESC	87	86	82	93	78	81	86	80	88	85	71	71
e) mouse adipocyte #1	89	86	88	79	85	87	88	88	69	86	73	69
f) mouse adipocyte #2	86	84	82	81	79	81	89	83	76	83	68	69
g) mouse adipocyte #3	89	83	85	80	83	85	88	87	68	85	69	72
h) mouse adipocyte #4	84	84	84	85	85	85	88	87	72	86	74	73
i) mouse fibroblast #1	85	81	79	91	78	79	83	78	90	81	64	62
j) mouse fibroblast #2	88	83	83	81	77	77	78	82	68	85	66	64
k) mouse hepatocyte #1	63	51	70	67	64	58	64	61	55	66	71	71
l) mouse hepatocyte #2	73	74	81	85	76	70	68	75	72	83	77	82

Figure 5.21: AUC-ROC and AUC-PR for CRUP and REPTILE. We trained our classifier and the REPTILE method on 12 samples from different cell types and species, and cell type-specific training sets with FANTOM5-based enhancers. For each classifier, we make predictions on all 12 test sets. Heatmaps summarize results for **a)** the area under the ROC-curve (AUC-ROC) and **b)** the area under the precision-recall curve (AUC-PR) for our classifier and REPTILE. Row labels correspond to training set and column labels to test set origin. All performance values are multiplied with 100, and are therefore within range of $[0, 100]$.

As can be seen, the AUC-ROC results are very high and stable for all the REPTILE classifiers ($\in [0.93, 0.99]$). The lowest results are achieved for the classifiers which were trained or tested on the mouse hepatocyte data. However, since the training set is unbalanced, the precision-recall curve is the more reliable measure (more details on AUC-ROC and AUC-PR in Section 3.3.4). The AUC-PR performances are less stable between different training and test set origins and vary between 0.51 and 0.93. Also here, the lowest performances appear for training or testing on mouse hepatocytes, with a range of $[0.51, 0.85]$ for training set origin and $[0.62, 0.82]$ for training set origin. In general, training and test set from the same data origin have similar AUC-PR results. The diagonal of the heatmap, which represents the performance of training and testing within a sample, shows in nearly half of the cases (5/12) the best result. In our method, which seems much more dependent on the test set alone, this only holds in 2/12 cases. A possible reason are the different normalization techniques. While we apply a quantile normalization towards the training data, REPTILE does not offer a built-in normalization technique but recommends a RPM normalization (see Section 4.5.2). This may not be enough to make the cutoff-based classifier transferable. Overall, our classifier shows a superior performance for training and prediction across cell types with an AUC-PR $\in [0.71, 0.95]$.

For a fair comparison, we also trained a REPTILE classifier close to its original setting in terms of feature and training enhancer choice as described in He *et al.* (2017) resulting in similar or slightly worse performances (Figure A7).

5.4 Summary and discussion

In this chapter, we validated our enhancer prediction method using a variety of criteria for the performance assessment, and compare it to two other methods. We observed that enhancer-associated HMs correspond to features which are of high importance in the decision making process in both mESC-based random forest classifiers. Moreover, the accessible region within an enhancer, and as such also the shape of the HM profiles, is represented in the measured feature importance. Applied on a carefully selected test set, we achieved very good performance results. In a genome-wide manner, our classifier showed a high spatial resolution as well as a reasonably low overlap with active promoters, which accounts for the most likely source of false positive predictions. Here, we also noticed the difficulties in distinguishing wrongly predicted active enhancers from actual promoter-proximal enhancers.

We demonstrated that our optimized classifier can be reliably applied across different cell types and species without the need of being re-trained. Furthermore, we showed that our classifier outperforms the prominent segmentation method ChromHMM, which is often used for enhancer prediction. Compared to the novel supervised approach REPTILE, we achieved similar performance results when training and predicting in the same condition. However, the spatial resolution of our most confidently predicted enhancers is higher, which could be due to the design of our feature set and the integration of several genomic bins at an enhancer region. In terms of transferability across conditions, our classifier performs better than REPTILE. A possible explanation could lie in the different normalization strategies. REPTILE as well as our classifier are cutoff-based methods, therefore relying on comparability of the input data between different samples. While REPTILE does not offer an integrated normalization technique for the HM experiments, we perform a quantile normalization and shift the distribution of new data towards the distribution of our training data.

6 Prediction of Dynamic Regulatory Units

In this chapter, we present a comprehensive framework to predict condition-specific regulatory units from histone modification and gene expression data called *CRUP*. The regulatory units consist of predicted enhancers that are dynamically changing between conditions of interest and their putative target genes which are located in the same topologically associating domain (TAD) and chosen based on similar activity patterns.

In the following, we analyze the advantages of differential enhancer (and regulatory unit) prediction and describe our three-step framework in more detail. Then, we demonstrate the application of our framework on a disease model of rheumatoid arthritis. Finally, we give a detailed description of the programmed pipeline which we made publicly available.

This chapter is based on results from a paper that I co-first authored with Verena Heinrich and that is available as a preprint (Ramisch *et al.*, 2018). The chapter covers parts of the second and third step of our framework *CRUP*, while the first step about modeling and validating an enhancer prediction method was extensively discussed in Chapters 4 and 5. Verena Heinrich and I designed the *CRUP* framework, discussed the applicational examples, and wrote the paper together with Martin Vingron. I did the motif enrichment analysis and wrote the code for training and predicting enhancers. Verena Heinrich assembled and revised the code for publication, as it can be found on github, and created the code for the second and third part of the *CRUP* framework (i.e., the differential enhancer calling and enhancer-gene matching).

6.1 Motivation

The prediction of enhancers in a genome-wide manner based on epigenomic data is an important task since enhancers play a big role in fully understanding the regulation of genes, and their misregulation is often related to complex diseases (Wang *et al.*, 2018). However, in many applications it is of interest to exclude ubiquitously active enhancers from the analysis and concentrate only on the set of cell type-specific enhancers and their corresponding target genes. This also holds true for the comparison of different conditions or states. To predict enhancers associated to a disease, it is necessary to find changes in enhancer activity between the healthy and the diseased state since both, the gain or the loss of activity, can be causative.

Similarly, to find disease-related or cell type-specific TFs, the motif search becomes easier upon excluding uniformly active enhancers by comparing different states. A good example here is the motif analysis we performed in Section 5.1.7 on the genome-wide predicted enhancers in mESC. We were able to find TFs associated to pluripotency as well as to embryonic development (e.g. *OCT4*, *TFAP2B*). However, without prior knowledge about these factors it would be difficult to interpret their possible role since they are not necessarily specific for embryonic stem cells. In Ramisch *et al.* (2018), we expanded the analysis and used our pre-trained classifier to additionally predict enhancers in an mESC sample which was treated with retinoic acid (RA) and therefore pushed into differentiation. Then, we used the second part of our framework to find a set of differentially expressed enhancers between the pluripotent and the differentiated state, and cluster them into two sets according to their activity level. For the motif analysis, we used the union of both clusters for the estimation of the background model. Some of the TFs which are enriched in our predicted mESC enhancers actually show an even higher enrichment in the RA-induced enhancers, e.g., *TFAP2B* or *ZNF384*, as can be seen in Figure 6.1. The set of enhancers active in the pluripotent but not in the RA-induced state is, among others, enriched for *OCT4* (*POU5F1*) and *HNF1A/B*, which are part of the *signaling pathways regulating pluripotency of stem cells* defined by KEGG (Kanehisa *et al.*, 2017). The enhancers

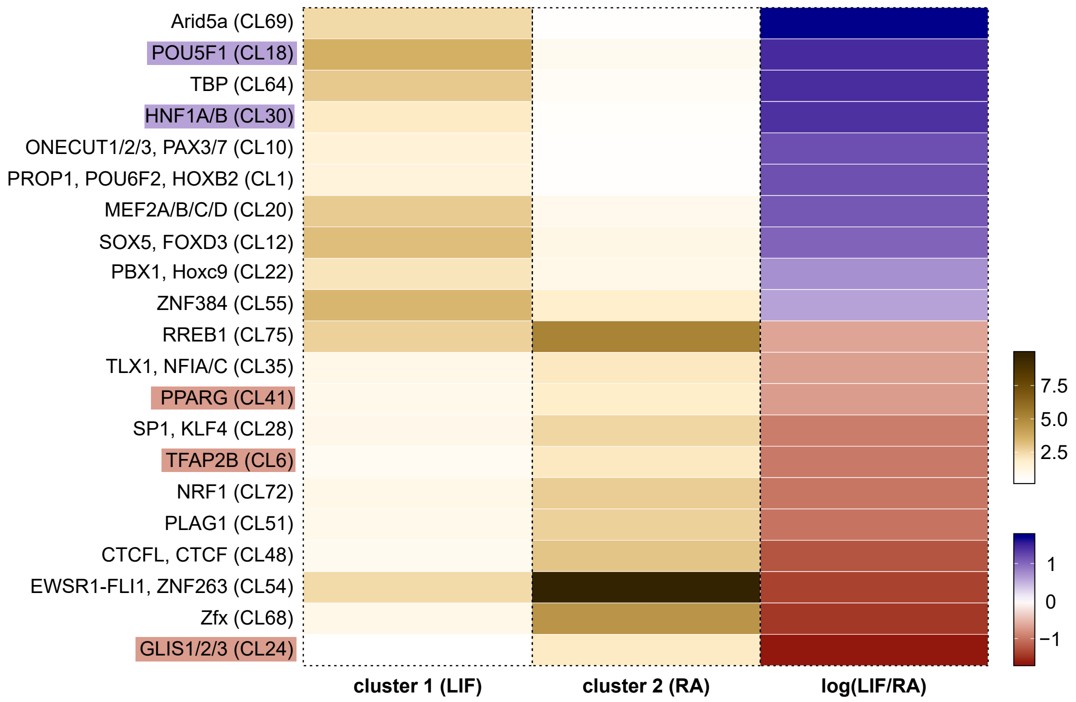


Figure 6.1: Differential motif analysis in mESC. Motif enrichment for differential enhancers in LIF and retinoic acid (RA) treated mESCs. Cluster 1 describes enhancers active in pluripotent but not differentiated state (LIF). Cluster 2 describes enhancer active in RA-induced differentiation state but not in pluripotent state (RA). Enrichment values for cluster 1 and 2, and the log-ratio between cluster enrichment for 21 JASPAR consensus motifs. The corresponding JASPAR motif cluster is indicated by ‘CL’, and a complete summary of the motif clusters can be found in Table B4. Consensus motifs shown have enrichment values ≥ 1 in one of the clusters and a differential enrichment ≥ 1. Log-fold values in red indicate motifs with higher motif enrichment in cluster 2, and in blue for cluster 1. Motif names marked in blue are enriched in cluster 1 and are part of signaling pathways regulating pluripotency in stem cells. TFs marked in red and enriched in cluster 2, form heterodimers with RA receptors (PPARG) or play roles in early development and differentiation.

exclusively active in the RA-induced cells show a differential enrichment for *PPARG*, which is known to form heterodimers with the RA-inducable *retinoid X receptor* (Mangelsdorf and Evans, 1995). Furthermore, *GLIS1/2/3*, potential regulators in the context of stem cell differentiation (Scoville *et al.*, 2017) and the RA-inducible activator of transcription *TFAP2B* (Luscher *et al.*, 1989)

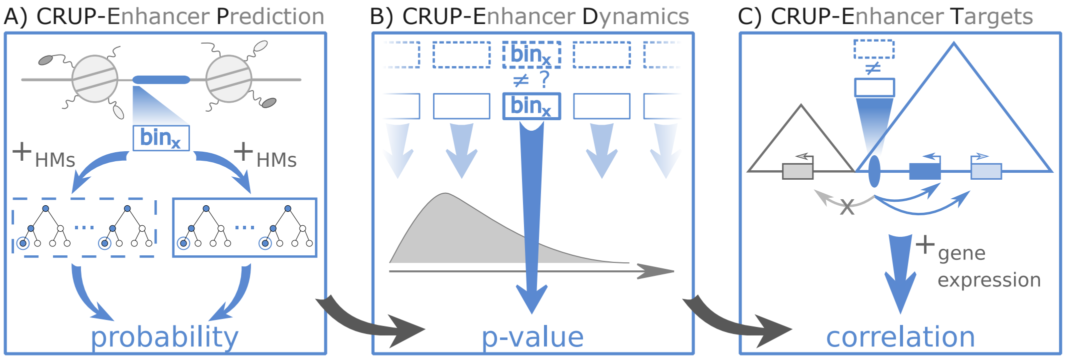


Figure 6.2: Workflow of CRUP.

show a higher enrichment in the RA-inducible enhancers. By applying the motif enrichment analysis to differential enhancers we gain knowledge about the condition-specific importance of TFs, and facilitate follow-up analyses.
Already mentioned, we are often not only interested in the location of an active enhancer but also in its target gene(s). Also here a possible application can be found in a disease-related scenario where we want to understand the cascade of misregulational events that causes a disease starting from the falsely activated or not activated enhancer.

6.2 Model description

CRUP, short for **C**ondition-specific **R**egulatory **U**nits **P**rediction), is a three-step framework depicted in Figure 6.2 to

A) **p**redict **e**nhancers with a pre-trained classifier ('CRUP-EP'),

B) find **d**ynamically changing **e**nhancers between multiple conditions ('CRUP-ED') and

C) match these **e**nhancers with putative **t**arget genes to build regulatory units ('CRUP-ET').

A) CRUP - Enhancer Prediction

The modeling, training and validation of our enhancer classifier is described in detail in Chapter 4. For our framework, we apply the pre-trained classifier to ChIP-seq data (H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K36me3, H3K9me3, control) of one or several cell types and conditions of interest. For each data sample, we get a list of annotated enhancers as well as a genome-wide distribution of predicted enhancer probabilities in 100 bp bins.

B) CRUP - Enhancer Dynamics

We compute differential enhancers between conditions using a non-parametric permutation test. The (binned) predicted enhancer probabilities are the input for the second part of CRUP. Several conditions can be included in the comparison, but each individual comparison is only between two conditions. Also, each condition can be represented by several samples or replicates.

Let C^1 and C^2 be two conditions. We store the predicted probabilities in two matrices, $A_{C^1} = (A_{xi})_{i \in C^1}$ and $A_{C^2} = (A_{xi})_{i \in C^2}$, where x indicates the genomic location (bin_x) and i the sample in condition C^1 or C^2. Since the number of samples is usually small, we use a non-parametric permutation test. Therefore, we shuffle each column (sample) of the two matrices individually and use them to compute an empirical distribution of the t-test statistic

$$T_x = \frac{\mu_{C^1} - \mu_{C^2} - w_0}{S_\triangle}.$$

Here, $\mu_{C^1} = \mu(A_{xC^1})$ and $\mu_{C^2} = \mu(A_{xC^2})$ are the condition-specific means for bin_x, w_0 set to 0.5 defines the minimum difference between them, and $S_\triangle$ is the pooled standard deviation defined as

$$S_\triangle^2 = \frac{(|C^1| - 1)\sigma_{C^1}^2 + (|C^2| - 1)\sigma_{C^2}^2}{|C^1| + |C^2| - 2} \cdot \Big(\frac{1}{|C^1|} + \frac{1}{|C^2|}\Big)$$

with condition-specific variances $\sigma_{C^1}^2 = \sigma^2(A_{xC^1})$ and $\sigma_{C^2}^2 = \sigma^2(A_{xC^2})$.

Based on this, we compute the empirical p-value $P_x = P_x(C^1, C^2)$ for each bin_x by counting the number of T_x that are higher then the true weighted difference T_x^{true} based on the unshuffled matrices A_{C^1} and A_{C^2}. Finally, by setting a threshold $P^* = 0.01$ on the empirical p-values P_x we find genomic bins (of size 100 bp) that show a significantly different enhancer probability between the two conditions and taking multiple samples into account when available.

Subsequently, we define condition-specific enhancers by clustering and summarizing differential bins based on their 'activity pattern'. Let the indicator function $T_x(C^1, C^2)$ be defined as

$$T_x(C^1, C^2) = \begin{cases} 1, & \text{if } P_{x-2:x+2}(C^1, C^2) \leq P^* \text{ and } (\mu_{C^1} - \mu_{C^2}) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

Then, $T_x(C^1, C^2) = 1$ denotes that bin_x is an active enhancer in condition C_1 but not in condition C_2. Note also that the indicator function $T_x(C^1, C^2)$ depends on the p-value assigned to bin_x, as well as the p-values of two additional bins upstream and downstream. Based on this, bin_x will be denoted as $\text{bin}_x^{\{T(C^1,C^2)=1,\ T(C^2,C^1)=0\}} = \text{bin}_x^{\{1,0\}}$ if it is significantly differential between conditions C^1 and C^2 with a higher condition-specific mean in condition C^1, $(\mu_{C^1} > \mu_{C^2})$, and as $\text{bin}_x^{\{0,1\}}$ if $(\mu_{C^1} < \mu_{C^2}$. In case of no differential activity, we write $\text{bin}_x^{\{0,0\}}$. With this, each bin_x can be allocated to a unique *'activity pattern'* of $\{1, 0\}$, $\{0, 1\}$ or $\{0, 0\}$ (see Figure 6.3 for an overview).

We can also apply this approach for a larger number of conditions. For three conditions, for example, the number of binary comparisons is $\binom{3}{2} = 3$ $((C^1, C^2), (C^1, C^3)$ and $(C^2, C^3))$ and the total number combinations of 'activity pattern' is $3^{\binom{3}{2}} - 1 = 26$. The pattern $\{0, 0, 0, 0, 0, 0\}$ can be excluded since it does not encode any differential information.

Finally, we define condition-specific enhancers by summarizing bins that have the same activity pattern and are located within a 2 kb distance. Here, the bin with the lowest empirical p-value is stored as peak. Note that regions with different activity patterns are summarized and labeled according to the lowest p-value.

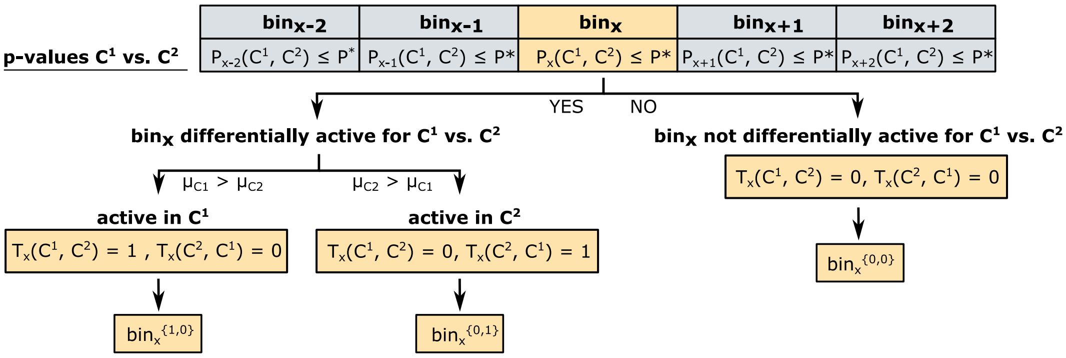


Figure 6.3: Strategy for the assignment of activity pattern based on two conditions. Comparison of enhancers in two conditions, C^1 and C^2. The activity pattern of bin$_x$ depends on the empirical p-values of bin$_x$ and two neighboring bins to both sides, the p-value cutoff P^*, and the group mean of C^1 and C^2 (μ_{C^1}, μ_{C^2}, respectively).

C) CRUP - Enhancer Targets

We predict putative target genes for the previously annotated differential enhancers based on their expression pattern across conditions. Therefore, we compute the expression counts per exon from the corresponding RNA-seq experiments of each sample in each condition and summarize them gene-wise (R function *summarizeOverlaps*, Lawrence *et al.* (2013)). Then, the counts per gene are variance stabilized (R function *vst*, Love *et al.* (2014)).

We elaborated in Section 2.4.8 that enhancer-promoter communication is mostly limited to domains of preferential chromatin interactions called topologically associating domains (TADs). Hence, to reduce the search space for potential target genes, we only consider enhancer-gene pairs in the same TAD. For each differential enhancer and each gene located in the same TAD, we compute the Pearson correlation between the predicted probability values and the normalized gene expression counts across all samples and conditions. Enhancer-gene pairs with a correlation $\geq$ 0.9 are considered as the final condition-specific regulatory units.

6.3 Description of experimental data

In the context of this chapter, we use mESC and mouse synovial fibroblast data sets, which we described in more detail the Appendix (Data Processing). To make comparisons between different conditions, we include a retinoic acid (RA)-induced mESC sample into our analysis as well as synovial fibroblast samples from mice affected with rheumatoid arthritis (Rh.A). An overview can be found in Table 6.1.

In order to reduce the search space for enhancer-gene interactions, we identify topologically associated domains (TADs) from Hi-C experiments in mESCs as explained in the Appendix (Data Processing).

Table 6.1: Summary of data samples and corresponding condition.

Abbreviation	Tissue/Cell type	Condition/Treatment
mESC (RA-induced)	ESC blastocyst	-LIF, +RA
mouse fibroblast #3 ('Mf06')	synovial fibroblast	Rh.A-like
mouse fibroblast #4 ('Mf08')	synovial fibroblast	Rh.A-like

6.4 Application to a disease model

We apply our framework to a complex disease study focusing on rheumatoid arthritis (Rh.A.), an autoimmune inflammatory disorder of the joints, which was part of the German Epigenome Program (DEEP, 2017). The goal is to find condition-specific regulatory units, i.e., enhancer gene-pairs either solely active in the healthy or in the diseased state, based on data from two healthy mice and two mice affected by rheumatoid arthritis.

Fist, we apply our pre-trained mESC classifier and predict genome-wide enhancer probabilities for all four samples. From these, we compute all differential 100 bp bins and summarize them according to the description in the previous section to 514 differential enhancers in total. Taking the TAD structure into account, we find 462 condition-specific regulatory units from which 60% (279) represent enhancer-gene pairs active in the rheumatoid arthritis affected mice (see Figure 6.4).

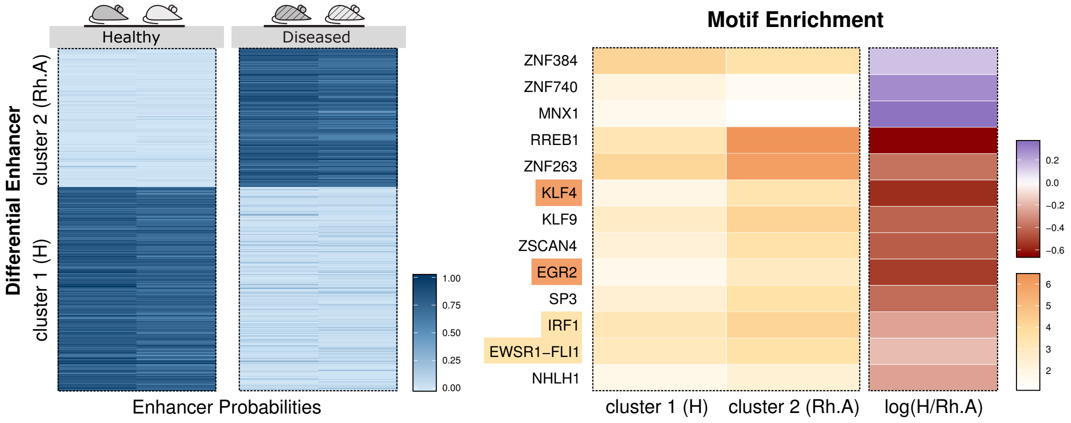


Figure 6.4: Clustering and motif analysis of disease-associated enhancers. Predicted differential enhancers for the rheumatoid arthritis disease model coloured according to their predicted probability and separated into two clusters based on their activity pattern. Cluster 1 (H) describes enhancers solely active in the healthy state and cluster 2 (Rh.A.) contains enhancers solely active in the disease state. Motif enrichment values for cluster 1 and 2, and the log-ratio between cluster enrichment for 13 JASPAR motifs with enrichment values ≥ 1 in one of the clusters and a differential enrichment ≥ 1. Log-fold values in red indicate motifs with higher motif enrichment in cluster 2, and in blue for cluster 1. Motif names marked in blue are enriched in cluster 1 and are part of signaling pathways regulating pluripotency in stem cells. TF motifs highlighted in orange, *KLF4* and *EGR2*, are directly associated to rheumatoid (Myouzen *et al.*, 2010; Luo *et al.*, 2016), while TF motifs marked in yellow, *IRF1* and *FLI1*, are connected to inflammatory conditions (Salem *et al.*, 2014) or regulate TFs associated with rheumatoid arthritis (Sato *et al.*, 2014).

We perform a motif analysis on all differential enhancers according to Section 4.4. Note that we use the binding profiles of all 579 TFs from the non-redundant JASPAR 2018 CORE vertebrate collection instead of the consensus motifs (Khan *et al.*, 2018). The four TFs *KLF4*, *EGR2*, *IRF1* and *FLI1* show a higher enrichment in the Rh.A. samples, see Figure 6.4, and are already known to be disease-associated (Myouzen *et al.*, 2010; Luo *et al.*, 2016; Salem *et al.*, 2014; Sato *et al.*, 2014).

An example region of several predicted regulatory units is shown in Figure 6.5. Here, a single differential enhancer which is active in the disease-affected mice is correlated to multiple putative target genes which all belong to the *CCR*-gene cluster and are part of the chemokine signaling pathway (KEGG). Chemokines

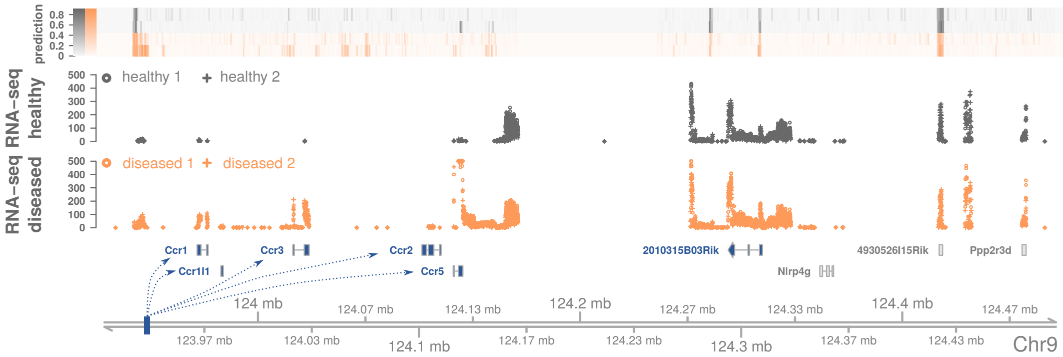


Figure 6.5: Example region containing differential regulatory units in a disease context. Predicted enhancer probabilities (‘*prediction*’) and RNA-seq data (raw counts, cut at a maximum of 500) for two healthy mice (gray) and two mice affected by rheumatiod arthritis (orange). Identified regulatory units consist of differential enhancer (blue bar) active in the diseased samples i six genes (highlighted in blue). Five of these genes are part of the *CCR*-gene cluster.

play a role in leukocyte recruitment to inflammation sites and are part of the Rh.A. pathogenesis (Zhang *et al.*, 2015). Additionally, the *IRF1* motif, which we found to be among the differentially enriched motifs, has a putative binding site in the involved differential enhancer. Already mentioned, *IRF1* is known to be connected to Rh.A (Salem *et al.*, 2014).

In this section we showed that our framework *CRUP* can be easily used to find promising candidates of condition-specific enhancer-gene pairs genome-wide and facilitate the search for disease causing misregulations.

6.5 Tool description

The three-step framework is realized in three independent functions: *CRUP - EP*, *CRUP - ED* and *CRUP - ET*. The function *CRUP - normalize* is added upstream for data preparation. Note that all four steps build upon each other following the order (1) *CRUP - normalize*, (2) *CRUP - EP*, (3) *CRUP - ED* and (4) *CRUP - ET*.

Function CRUP - normalize

The function computes the read counts for HM and control/ input ChIP-seq experiments in 100 bp bin. Then, an input-normalization is performed.

Function CRUP - EP

A pre-trained random-forest based classifier is applied to the input-normalized count distributions from the previous step. The input-normalized counts are, just before applying the classifier, quantile normalized to the distribution of the corresponding training samples to guarantee a good transferability. The output of this function is a list of putative enhancers as well as genome-wide predicted enhancer probabilities for each 100 bp.

Function CRUP - ED

Applying a non-parametric permutation test to the predicted probabilities from the previous step, this function identifies condition-specific or dynamically changing enhancers for several conditions (and samples).

Function CRUP - ET

The functions predicts condition-specific regulatory units from the predicted probabilities of the previously annotated condition-specific enhancers and RNA-seq data from the same conditions (and samples).

Availability

The code to run our tool *CRUP* is openly available at `https://github.com/VerenaHeinrich/CRUP`.

6.6 Summary and discussion

In this chapter, we described a 3-step framework to predict condition-specific regulatory units based on histone modification and gene expression data. In the first step, a pre-trained classifier is applied to HM ChIP-seq data of several user-defined conditions to make genome-wide enhancer predictions. Then, differentially active enhancers are assigned to their corresponding conditions using a non-parametric permutation test on the predicted enhancer probabilities. Since low quality data can have a strong influence on the individual enhancer predictions and therefore also on the permutation test performance, it is advisable to use several samples or replicates for each condition, if available. Note that the number of conditions is not limited since the test is performed in a pair-wise manner. Finally, condition-specific enhancers are linked to putative target genes located within the same topologically associating domain (TAD). To this end,the correlation between enhancer probabilities and normalized gene expression values across all conditions is computed and using a very high threshold the final list of condition-specific regulatory units is obtained. For all our analyses, we use TAD annotations from mouse embryonic stem cells (Rao *et al.*, 2002) and argue that the structure of TADs are (to a certain degree) invariant between cell types and conserved between related species (see also Section 2.4.8).

We applied our framework to a mouse study on rheumatoid arthritis (Rh.A) and were able to find $\sim$ 450 disease-associated regulatory units genome-wide. A motif analysis on the differentially active enhancers revealed an enrichment of several TFs known to play a role in rheumatoid arthritis. We also found an interesting genomic region where one differential enhancer was linked to several genes belonging to a gene cluster associated with the rheumatoid arthritis pathogenesis.

In summary, we introduced a freely available tool to collapse several layers of epigenetic data to a candidate list of condition-specific enhancer-gene pairs. These candidates could be used for a more comprehensive analysis of single genomic loci and therefore dramatically reduce the search space for condition-oriented research based on genome-wide data.

7 Conclusion and Discussion

In this work, we introduced a novel supervised enhancer prediction method that can locate active enhancers genome-wide in a chromosomal context. The design of our model and the feature set are based on prior knowledge on several enhancer features. We offer a pre-trained classifier which can be applied to new data without re-training, and subsequently demonstrate how it can be used as a first step in a comprehensive framework to predict condition-specific pairs of enhancers and putative target genes.

The machine-learning methods used for enhancer prediction can be categorized into unsupervised, semi-supervised and supervised approaches (see Chapter 3 for a general introduction of machine-learning). ChromHMM is a prominent unsupervised genome-segmentation tool often used for enhancer prediction, which can be applied to new data without prior knowledge. However, the emitted genome segmentation has to be interpreted and annotated by the user, which can lead to highly variable performance results as we demonstrated in 5.3. Our classifier obtained stable results when trained and tested in the same cell type and showed higher performances than ChromHMM in several different settings. The most common bottleneck for supervised enhancer prediction is the lack of prior knowledge. Only few available methods offer pre-trained classifiers which are transferable to experimental settings for which no training data exists. In Section 5.2, we showed that our classifier obtains good performance results across different cell types, tissues and species, and we chose the most reliable pre-trained classifier for further analyses. We also saw that our classification method is comparable to another supervised method for enhancer prediction, REPTILE, and is superior concerning transferability to new cell types or species (see Section 5.3). Here, we argued that next to differences in modeling and training set choice especially our built-in quantile normalization seems to be responsible for the superior performance. Since both classifiers are cutoff-based, it is very important to account for feature differences between conditions, for example due to varying sequencing depth. In the past decades, several enhancer properties have been observed on individual validated example regions. We discussed the majority of these properties extensively in Section 2.4. Recent innovations in sequencing techniques such as high-throughput sequencing accelerated the research in this direction since it lead to the production of comprehensive genome-wide data collections covering many different cell types and conditions. A selection of experimental techniques used to study gene regulation can be found in Section 2.3. Currently, many enhancer properties are exploited for computational enhancer prediction.

Our classifier, described in detail in Chapter 4, is based on histone modifications as features, and uses accessibility and bidirectional transcription to define active enhancers for training. We showed in Chapter 4 that this selection of enhancer properties enabled reliable results for genome-wide enhancer prediction.

However, looking at several observed enhancer properties in more detail, it became apparent that no property alone covers all active enhancers in one condition. Krebs *et al.* (2011) showed that the histone acetyltransferase p300 is not bound at a set of active enhancers in two human cell lines, even though it is vastly used as an approximation for enhancer activity, as for example by the REPTILE method. Pradeepa *et al.* (2016) found active enhancers in mouse embryonic stem cells that were not marked by H3K27ac but showed enrichment of the histone modification H3K122ac instead. Sequence conservation and DNA methylation are other known examples for features that are present at some enhancers but not others. Furthermore, it was recently claimed that bidirectional transcription is not exclusively found at active enhancers, but more broadly coincides with accessible chromatin (Young *et al.*, 2017). Thus, we have to bear in mind the possible biases that every enhancer prediction method is introducing through the underlying enhancer properties of the feature and training criteria. We also saw this in Section 5.3 during the comparison of our classifier to the supervised approach REPTILE which is based on a very similar HM feature set and p300-bound training enhancers. Here, our analysis indicated that training enhancers defined by bidirectional transcription resulted in a higher fraction of promoter-proximal predictions than p300-bound training enhancers. Interestingly, a class of enhancer-like promoters, also called 'Epromoters', were discovered recently and shown to regulate the expression of distal genes (see Review by Dao and Spicuglia (2018)).

All these observations raise the question, if the definition of one broad class of active enhancers is possible or even desirable in the field of enhancer prediction. Furthermore, they illustrate the difficulties to unite the original definition of enhancers, which is based on their functionality to activate transcription of a target gene, with the multiple convenient and necessary approaches to characterize enhancers by certain features.

Recently, Arnold *et al.* (2014) put forward a method called STARR-seq, short for 'self-transcribing active regulatory region sequencing', to map enhancer-activity for millions of genomic regions using quantitative enhancer assays. The method makes use of reporter constructs containing a minimal promoter and a downstream candidate sequence, and is based on the idea that an enhancer will initiate its own transcription. Hence, STARR-seq offers a possibility to detect thousands of regions in parallel which are able to drive transcription in the described episomal setting. However, it was found that mapped back into the original chromosomal setting, a subset of the putative enhancers are not active and carry repressive HM marks such as H3K27me3. Another observation was, that the enhancer activity depends on the minimal promoter used in the constructs hinting towards a necessary compatibility between an enhancer and its target promoter. It seems that the right chromatin environment and locus topology are crucial for the final activity status of a genomic region which is 'in theory' able to act as an enhancer. Thus, besides direct (functional) and indirect (property-based) characterizations of enhancers, also differences in episomal and chromosomal settings have to be taken into account for a successful enhancer prediction. In the future, this could be realized by combining known enhancer features measured in the original chromosomal context with larger sets of functional enhancers detected with STARR-seq or similar approaches. Unfortunately, such enhancer sets are still rare and some systematic errors in the quite recent STARR-seq method have only been resolved recently (Muerdter *et al.*, 2017).

By integrating multiple properties of enhancers into our classification model, we tried to compensate for the lack of functionally defined enhancers in the cell types we worked with. Additionally, we continued our work towards an application-oriented direction. In Chapter 6 we describe a comprehensive framework which is built on our pre-trained classifier and predicts condition-specific regulatory units, i.e., enhancer-gene pairs changing dynamically across conditions. Limiting enhancers to specific conditions has the advantage of introducing another layer of information that can be used to filter out weak predictions. In Section 6.1, the motif analysis of enhancers with differential active patterns between two states lead to more meaningful and specific enrichment results than the same analysis on the full set of predicted enhancers in one of the conditions. Hence, an application-oriented approach can help to better understand gene regulation mechanisms. In the last step of our framework, we use gene expression data to match differential enhancers with putative target genes. By reducing the set of possible enhancer targets to genes located in the same topologically associating domain, we integrate another layer of chromosomal information and offer a manageable candidate list of regulatory units that can facilitate further loci-based analyses.

Bibliography

Agalioti, T., Chen, G., and Thanos, D. (2002). Deciphering the transcriptional histone acetylation code for a human gene. *Cell*, **111**(3), 381–92.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular biology of the cell*. Garland Science.

Amos, S. (2008). When Training and Test Sets Are Different: Characterizing Learning Transfer. In *Dataset Shift Mach. Learn.*, pages 2–28. The MIT Press.

Andersson, R. *et al.* (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, **507**(7493), 455–461.

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., Lau, N. C., and Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat. Genet.*, **46**(7), 685–692.

Arrigoni, L., Richter, A. S., Betancourt, E., Bruder, K., Diehl, S., Manke, T., and Bnisch, U. (2016). Standardizing chromatin research: a simple and universal method for chip-seq. *Nucleic acids research*, **44**, e67.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, **27**(2), 299–308.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, **33**(3), 729–740.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, **129**(4), 823–837.

Bell, A. C., West, A. G., and Felsenfeld, G. (2001). Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science*, **291**(5503), 447–450.

Bellen, H. J., O'Kane, C. J., Wilson, C., Grossniklaus, U., Pearson, R. K., and Gehring, W. J. (1989). P-element-mediated enhancer detection: a versatile method to study development in Drosophila. *Genes Dev.*, **3**(9), 1288–1300.

Bembom, O. (2018). *seqLogo: Sequence logos for DNA sequence alignments*. R package version 1.48.0.

Bernstein, B. E. *et al.* (2005). Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell*, **120**(2), 169–181.

Bernstein, B. E. *et al.* (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, **125**(2), 315–326.

Bernstein, B. E. *et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**(10), 1045–1048.

Blow, M. J. *et al.* (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, **42**(9), 806–810.

Bolstad, B. (2018). *preprocessCore: A collection of pre-processing functions*. R package version 1.42.0.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A., and Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, **171**(3), 557–572.e24.

Boyer, L. A. *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**(6), 947–956.

Boyer, L. A. *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**(7091), 349–353.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**(2), 311–322.

Breiman, L. (2001). Random Forests. *Mach. Learn.*, **45**(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall.

Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y., and Allis, C. (1996). Tetrahymena Histone Acetyltransferase A: A Homolog to Yeast Gcn5p Linking Histone Acetylation to Gene Activation. *Cell*, **84**(6), 843–851.

Buecker, C. and Wysocka, J. (2012). Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet.*, **28**(6), 276–84.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.*, **109**, 21.29.1–9.

Bulger, M., Schübeler, D., Bender, M. A., Hamilton, J., Farrell, C. M., Hardison, R. C., and Groudine, M. (2003). A complex chromatin landscape revealed by patterns of nuclease sensitivity and histone modification within the mouse beta-globin locus. *Mol. Cell. Biol.*, **23**(15), 5234–44.

Calo, E. and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, **49**(5), 825–837.

Carninci, P. *et al.* (2005). The transcriptional landscape of the mammalian genome. *Science*, **309**(5740), 1559–1563.

Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**(13), e119.

Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked sites. *Genetics*, **190**(1), 5–22.

Chen, D., Ma, H., Hong, H., Koh, S. S., Huang, S. M., Schurter, B. T., Aswad, D. W., and Stallcup, M. R. (1999). Regulation of transcription by a protein methyltransferase. *Science*, **284**(5423), 2174–2177.

Chen, X. *et al.* (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell*, **133**(6), 1106–1117.

Cheng, J. *et al.* (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**(5725), 1149–1154.

Clark, M. B. *et al.* (2011). The reality of pervasive transcription. *PLoS Biol.*, **9**(7), e1000625; discussion e1001102.

Corces, M. R. *et al.* (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods.*, **14**(10), 959–962.

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**(1), 1–13.

Crawford, G. E. *et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.*, **16**(1), 123–131.

Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**(4), 292–301.

Dao, L. T. M. and Spicuglia, S. (2018). Transcriptional regulation by promoters with enhancer function. *Transcription*, **9**(5), 307–314.

Daugherty, A. C., Yeo, R. W., Buenrostro, J. D., Greenleaf, W. J., Kundaje, A., and Brunet, A. (2017). Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res.*, **27**(12), 2096–2107.

Davie, K., Jacobs, J., Atkins, M., Potier, D., Christiaens, V., Halder, G., and Aerts, S. (2015). Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet.*, **11**(2), e1004994.

De Santa, F. *et al.* (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol.*, **8**(5), e1000384.

DEEP (2012–2017). The German epigenome programme. http://www.deutsches-epigenom-programm.de.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–380.

Dobin, A., Carrie, A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1), 15–21.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. Wiley.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., and Lieberman Aiden, E. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, **3**(1), 95–98.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Ernst, J. *et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345), 43–49.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**(8), 817–25.

Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**(3), 215–216.

Ernst, J. and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**(12), 2478–2492.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**(8), 861–874.

Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J. R., and Bickmore, W. A. (2008). Recruitment to the Nuclear Periphery Can Alter Expression of Genes in Human Cells. *PLoS Genet.*, **4**(3), e1000039.

Firpi, H. A., Ucar, D., and Tan, K. (2010). Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**(13), 1579–1586.

Fu, S., Wang, Q., Moore, J. E., Purcaro, M. J., Pratt, H. E., Fan, K., Gu, C., Jiang, C., Zhu, R., Kundaje, A., Lu, A., and Weng, Z. (2018). Differential analysis of chromatin accessibility and histone modifications for predicting mouse developmental enhancers. *Nucleic Acids Res.*, **46**(21), 11184–11201.

Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell*, **33**(3), 717–728.

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**(6), 877–885.

Godoy, P. *et al.* (2013). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Arch Toxicol.*, **87**(8), 1315–530.

Hamada, H. (1986). Random isolation of gene activator elements from the human genome. *Mol. Cell. Biol.*, **6**(12), 4185–94.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

He, Y., , *et al.* (2017). Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc. Natl. Acad. Sci. U. S. A.*, **114**(9), E1633–E1640.

Hebbes, T. R., Thorne, A. W., and Crane-Robinson, C. (1988). A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J.*, **7**(5), 1395–402.

Heintzman, N. D. *et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**(3), 311–318.

Heintzman, N. D. *et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**(7243), 108–112.

Hesselberth, J. R. *et al.* (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**(4), 283–289.

Hinrichs, A. S. *et al.* (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, **34**, D590–598.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, **155**(4), 934–947.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**(5), 473–476.

Holliday, R. and Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. *Science*, **187**(4173), 226–232.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA*, **8**(1).

Huska, M. R., Ramisch, A., Vingron, M., and Marsico, A. (2016). Predicting enhancers using a small subset of high confidence examples and co-training. *PeerJ Preprints*, **4**, e2407v1.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer Texts in Statistics. Springer New York.

Javierre, B. M. *et al.* (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, **167**(5), 1369–1384.e19.

Johnson, J. M., Edwards, S., Shoemaker, D., and Schadt, E. E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, **21**(2), 93–102.

Johnson, K. D., Grass, J. A., Park, C., Im, H., Choi, K., and Bresnick, E. H. (2003). Highly restricted localization of RNA polymerase II within a locus control region of a tissue-specific chromatin domain. *Mol. Cell. Biol.*, **23**(18), 6484–6493.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**(7), 484–492.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E., Birney, E., and Furlong, E. (2012). A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. *Cell*, **148**(3), 473–486.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

Kapranov, P. *et al.* (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, **316**(5830), 1484–1488.

Keffer, J., Probert, L., Cazlaris, H., Georgopoulos, S., Kaslaris, E., Kioussis, D., and Kollias, G. (1991). Transgenic mice expressing human tumour necrosis factor: A predictive genetic model of arthritis. *EMBO J.*, **10**(13), 4025–4031.

Kellis, M. *et al.* (2014). Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, **111**(17), 6131–6138.

Kersey, P. J. *et al.* (2018). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.*, **46**(D1), D802–D808.

Khan, A. *et al.* (2018). Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**(D1), D260–D266.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature*, **436**(7052), 876–880.

Kim, T.-K. *et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**(7295), 182–187.

Kinkley, S., Helmuth, J., Polansky, J. K., Dunkel, I., Gasparoni, G., Froehler, S., Chen, W., Walter, J., Hamann, A., and Chung, H. (2016). reChIP-seq reveals widespread bivalency of H3K4me3 and H3K27me3 in CD4+ memory T cells. *Nature Communications*, **7**.

Kleftogiannis, D., Kalnis, P., and Bajic, V. B. (2016). Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinform.*, **17**(6), 967–979.

Knight, P. and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA J Numer Anal*, **33**(3), 1029–1047.

Koch, F. *et al.* (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.*, **18**(8), 956–963.

Kopp, W. (2017). *motifcounter: R package for analysing TFBSs in DNA sequences*. R package version 1.5.4.

Kopp, W. and Vingron, M. (2017). An improved compound Poisson model for the number of motif hits in DNA sequences. *Bioinformatics*, **33**(24), 3929–3937.

Koski, T. (2001). *Hidden Markov models for bioinformatics*. Kluwer Academic Publishers.

Krebs, A. R., Karmodiya, K., Lindahl-Allen, M., Struhl, K., and Tora, L. (2011). SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers. *Mol. Cell*, **44**(3), 410–423.

Kulkarni, M. M. and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development*, **130**(26), 6569–6575.

Lam, M. T. Y. *et al.* (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature*, **498**(7455), 511–515.

Lasserre, J., Chung, H.-R., and Vingron, M. (2013). Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks. *PLoS Comput. Biol.*, **9**(9), e1003168.

Latinne, P., Latinne, P., Saerens, M., and Decaestecker, C. (2001). Adjusting the Outputs of a Classifier to New a Priori Probabilities May Significantly Improve Classification Accuracy: Evidence from a Multi-Class Problem in Remote Sensing. *NEURAL Comput.*, **14**, 14–21.

Lawrence, M., Huber, W., Pags, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, **9**.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*, **7**(9), 709–715.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Curr. Biol.*, **20**(17), R754–763.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, W. *et al.* (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, **498**(7455), 516–520.

Liang, G., Lin, J. C. Y., Wei, V., Yoo, C., Cheng, J. C., Nguyen, C. T., Weisenberger, D. J., Egger, G., Takai, D., Gonzales, F. A., and Jones, P. A. (2004). Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc. Natl. Acad. Sci.*, **101**(19), 7357–7362.

Lieberman-Aiden, E. *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950), 289–293.

Lim, L. W. K., Chung, H. H., Chong, Y. L., and Lee, N. K. (2018). A survey of recently emerged genome-wide computational enhancer predictor tools. *Comput. Biol. Chem.*, **74**, 132–141.

Lin, Q., Lu, J., Yanagisawa, H., Webb, R., Lyons, G. E., Richardson, J. A., and Olson, E. N. (1998). Requirement of the MADS-box transcription factor MEF2C for vascular development. *Development*, **125**(22), 4565–4574.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, **133**(3), 523–536.

Litt, M. D., Simpson, M., Gaszner, M., Allis, C. D., and Felsenfeld, G. (2001). Correlation Between Histone Lysine Methylation and Developmental Changes at the Chicken beta -Globin Locus. *Science*, **293**(5539), 2453–2455.

Long, H. K., Prescott, S. L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, **167**(5), 1170–1187.

Louppe, G. (2014). Understanding Random Forests: From Theory to Practice. *eprint arXiv:1407.7502*.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Luo, X., Chen, J., Ruan, J., Chen, Y., Mo, X., Xie, J., and Lv, G. (2016). Krüppel-Like Factor 4 Is a Regulator of Proinflammatory Signaling in Fibroblast-Like Synoviocytes through Increased IL-6 Expression. *Mediators Inflamm.*, **2016**, 1062586.

Luscher, B., Mitchell, P. J., Williams, T., and Tjian, R. (1989). Regulation of transcription factor AP-2 by the morphogen retinoic acid and by second messengers. *Genes Dev.*, **3**(10), 1507–1517.

Mammana, A. and Chung, H.-R. (2015). Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.*, **16**(1), 151.

Mammana, A. and Helmuth, J. (2016). *bamsignals: Extract read count signals from bam files*. R package version 1.12.1.

Mangelsdorf, D. J. and Evans, R. M. (1995). The RXR heterodimers and orphan receptors. *Cell*, **83**(6), 841–850.

May, D. *et al.* (2012). Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.*, **44**(1), 89–93.

Melgar, M. F., Collins, F. S., and Sethupathy, P. (2011). Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.*, **12**(11), R113.

Merika, M. and Thanos, D. (2001). Enhanceosomes. *Curr. Opin. Genet. Dev.*, **11**(2), 205–208.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**(7), 621–628.

Moser, M., Pscherer, A., Roth, C., Becker, J., Mucher, G., Zerres, K., Dixkens, C., Weis, J., Guay-Woodford, L., Buettner, R., and Fassler, R. (1997). Enhanced apoptotic cell death of renal epithelial cells in mice lacking transcription factor AP-2beta. *Genes Dev.*, **11**(15), 1938–1948.

Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G. L., and Sartorelli, V. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell*, **51**(5), 606–617.

Muerdter, F. *et al.* (2017). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods*, **15**(2), 141–149.

Myouzen, K. *et al.* (2010). Regulatory polymorphisms in EGR2 are associated with susceptibility to systemic lupus erythematosus. *Hum. Mol. Genet.*, **19**(11), 2313–2320.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, **320**(5881), 1344–1349.

Natoli, G. and Andrau, J.-C. (2012). Noncoding Transcription at Enhancers: General Principles and Functional Models. *Annu. Rev. Genet.*, **46**(1), 1–19.

Ng, F. S., Schütte, J., Ruau, D., Diamanti, E., Hannah, R., Kinston, S. J., and Göttgens, B. (2014). Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells. *Nucleic Acids Res.*, **42**(22), 13513–13524.

Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. (2003). Scanning Human Gene Deserts for Long-Range Enhancers. *Science*, **302**(5644), 413.

Nora, E. P. *et al.* (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398), 381–385.

Novo, C. L. *et al.* (2018). Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. *Cell Rep.*, **22**(10), 2615–2627.

Ørom, U. A. *et al.* (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**(1), 46–58.

Pease, S., Braghetta, P., Gearing, D., Grail, D., and Williams, R. L. (1990). Isolation of embryonic stem (ES) cells in media supplemented with recombinant leukemia inhibitory factor (Lif). *Nature*, **141**(2), 344–352.

Pennacchio, L. A. *et al.* (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**(7118), 499–502.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nat. Rev. Genet.*, **14**(4), 288–295.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, **17**(5), 556–565.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E. M., Couronne, O., and Pennacchio, L. A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.*, **16**(7), 855–863.

Pradeepa, M. M., Grimes, G. R., Kumar, Y., Olley, G., Taylor, G. C. A., Schneider, R., and Bickmore, W. A. (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.*, **48**(6), 681–686.

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. *Nature*, **322**(6081), 697–701.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**(7333), 279–283.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., and Ren, B. (2013). RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Comput. Biol.*, **9**(3), e1002968.

Ramírez, F. *et al.* (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**(W1), W160–W165.

Ramisch, A. *et al.* (2018). CRUP: A comprehensive framework to predict condition-specific regulatory units. *bioRxiv*, page 501601.

Rao, S., Huntley, M., Durand, N., Stamenova, E., Bochkov, I., Robinson, J., Sanborn, A., Machol, I., Omer, A., and Lander, E. (2002). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Rea, S. *et al.* (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, **406**(6796), 593–599.

Riggs, A. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenet. Genome Res.*, **14**(1), 9–25.

Roh, T.-Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.*, **19**(5), 542–552.

Roh, T.-Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. *Proc. Natl. Acad. Sci.*, **103**(43), 15782–15787.

Rowley, M. J. and Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nat. Rev. Genet.*, **19**(12), 789–800.

Sabo, P. J. *et al.* (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods*, **3**(7), 511–518.

Salem, S., Gao, C., Li, A., Wang, H., Nguyen-Yamamoto, L., Goltzman, D., Henderson, J. E., and Gros, P. (2014). A novel role for interferon regulatory factor 1 (IRF1) in regulation of bone metabolism. *J. Cell. Mol. Med.*, **18**(8), 1588–98.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.*, **3**(3), 210–229.

Sato, S., Lennard Richard, M., Brandon, D., Jones Buie, J. N., Oates, J. C., Gilkeson, G. S., and Zhang, X. K. (2014). A critical role of the transcription factor fli-1 in murine lupus development by regulation of interleukin-6 expression. *Arthritis Rheumatol. (Hoboken, N.J.)*, **66**(12), 3436–44.

Schmidl, C., Klug, M., Boeld, T. J., Andreesen, R., Hoffmann, P., Edinger, M., and Rehli, M. (2009). Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res.*, **19**(7), 1165–1174.

Schmidt, D. *et al.* (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**(5981), 1036–1040.

Schmidt, F. o. (2016). Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, **45**(1).

Schneider, R., Bannister, A. J., Myers, F. A., Thorne, A. W., Crane-Robinson, C., and Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat. Cell Biol.*, **6**(1), 73–77.

Schübeler, D. *et al.* (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.*, **18**(11), 1263–1271.

Scoville, D. W., Kang, H. S., and Jetten, A. M. (2017). GLIS1-3: emerging roles in reprogramming, stem and progenitor cell differentiation and maintenance. *Stem Cell Investig.*, **4**, 80.

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., and Zhou, X. J. (2016). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research*, **44**(7), e70.

Shiraki, T. *et al.* (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, **100**(26), 15776–15781.

Siepel, A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**(8), 1034–50.

Smith, A. G., Heath, J. K., Donaldson, D. D., Wong, G. G., Moreau, J., Stahl, M., and Rogers, D. (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature*, **336**(6200), 688–690.

Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**(3), 204–220.

Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**(9), 613–626.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K., Schübeler, D., and Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**(7378), 490–495.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, **403**(6765), 41–45.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, **14**(2), 103–105.

Stunnenberg, H. G., The International Human Epigenome Consortium, and Hirst, M. (2016). The international human epigenome consortium: A blueprint for scientific collaboration and discovery. *Nat Biotechnol.*, **167**(5), 1145–1149.

Sun, F.-L. and Elgin, S. C. (1999). Putting Boundaries on Silence. *Cell*, **99**(5), 459–462.

Sun, W. *et al.* (2016). Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell*, **167**(5), 1385–1397.e11.

Sur, I. and Taipale, J. (2016). The role of enhancers in cancer. *Nat. Rev. Cancer*, **16**(8), 483–493.

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.*, **7**(3), 542–561.

Thanos, D. and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell*, **83**(7), 1091–1100.

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**(7146), 799–816.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Mol. Cell*, **10**(6), 1453–1465.

van Bakel, H., Nislow, C., Blencowe, B. J., and Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biol.*, **8**(5), e1000371.

van Holde, K. E. (1989). *Chromatin*. Springer Series in Molecular Biology. Springer New York, New York, NY.

Visel, A. *et al.* (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**(7231), 854–858.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**(Database), D88–D92.

Wang, D. *et al.* (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**(7351), 390–404.

Wang, Z. *et al.* (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**(7), 897–903.

Wang, Z., Zhang, Q., Zhang, W., Lin, J.-R., Cai, Y., Mitra, J., and Zhang, Z. D. (2018). HEDD: Human Enhancer Disease Database. *Nucleic Acids Res.*, **46**(D1), D113–D120.

Weber, F., de Villiers, J., and Schaffner, W. (1984). An SV40 enhancer trap incorporates exogenous enhancers or generates enhancers from its own sequences. *Cell*, **36**(4), 983–992.

Wehmeyer, C. *et al.* (2016). Sclerostin inhibition promotes TNF-dependent inflammatory joint destruction. *Science Translational Medicine*, **8**(330), 330ra35.

Whitaker, J. W., Nguyen, T. T., Zhu, Y., Wildberg, A., and Wang, W. (2015). Computational schemes for the prediction and annotation of enhancers from epigenomic assays. *Methods*, **72**, 86–94.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**(2), 307–319.

Woolfe, A. *et al.* (2004). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biol.*, **3**(1), e7.

Wu, C. (1980). The 5 ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, **286**(5776), 854–860.

Wysoker, A., Tibbetts, K., and Fennell, T. (2013). Picard tools. http://picard.sourceforge.net.

Young, R. S., Kumar, Y., Bickmore, W. A., and Taylor, M. S. (2017). Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol.*, **18**, 242.

Zalzman, M. *et al.* (2010). Zscan4 regulates telomere elongation and genomic stability in ES cells. *Nature*, **464**(7290), 858–863.

Zhang, L., Yu, M., Deng, J., Lv, X., Liu, J., Xiao, Y., Yang, W., Zhang, Y., and Li, C. (2015). Chemokine Signaling Pathway Involved in CCL2 Expression in Patients with Rheumatoid Arthritis. *Yonsei Med. J.*, **56**(4), 1134–1142.

Zhang, P., Andrianakos, R., Yang, Y., Liu, C., and Lu, W. (2010). Kruppel-like factor 4 (Klf4) prevents embryonic stem (ES) cell differentiation by regulating Nanog gene expression. *J. Biol. Chem.*, **285**(12), 9180–9189.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**(9), R137.

Ziller, M. J. *et al.* (2011). Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet.*, **7**(12), e1002389.

Ziller, M. J. *et al.* (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**(7463), 477–481.

List of Abbreviations

ATAC	Assay for transposase accessible chromatin
AUC	Area under the curve
AUC-PR	Area under the precision-recall curve
AUC-ROC	Area under the ROC curve
bp	Base pair
CAGE	Cap analysis gene expression
CGI	CpG island
ChIP	Chromatin immunoprecipitation
DHS	DNase I hypersensitive site
DMR	Differentially methylated region
EM	Expectation maximization
ESC	Embryonic stem cell
FAIRE	Formaldehyde-assisted isolation of regulatory elements
FANTOM	Functional annotation of the mouse/mammalian genome
FN	False negative
FP	False positive
FPKM	Fragments per kilobase million
FPR	False positive rate
HAT	Catalytic histone acetyltransferase
HM	Histone modification
HMM	Hidden Markov model
kp	Kilo base
LIF	leukemia inhibitory factor
mESC	Mouse embryonic stem cell

mESC_{REP}	Mouse embryonic stem cell data used by He *et al.* (2017)
NGS	Next-generation sequencing
PFM	Position frequency matrix
PREC	precision
Rh.A.	Rheumatoid arthritis
ROC	Receiver operating characteristics
SE	Super enhancer
TAD	Topologically associating domain
TN	True negative
TNR	True negative rate
TF	Transcription factor
TFBS	Transcription factor binding site
TP	True positive
TPM	Tags per million
TPR	True positive rate
TSS	Transcription start site

List of Figures

List of Tables

Appendices

Appendix A
Supplementary Figures

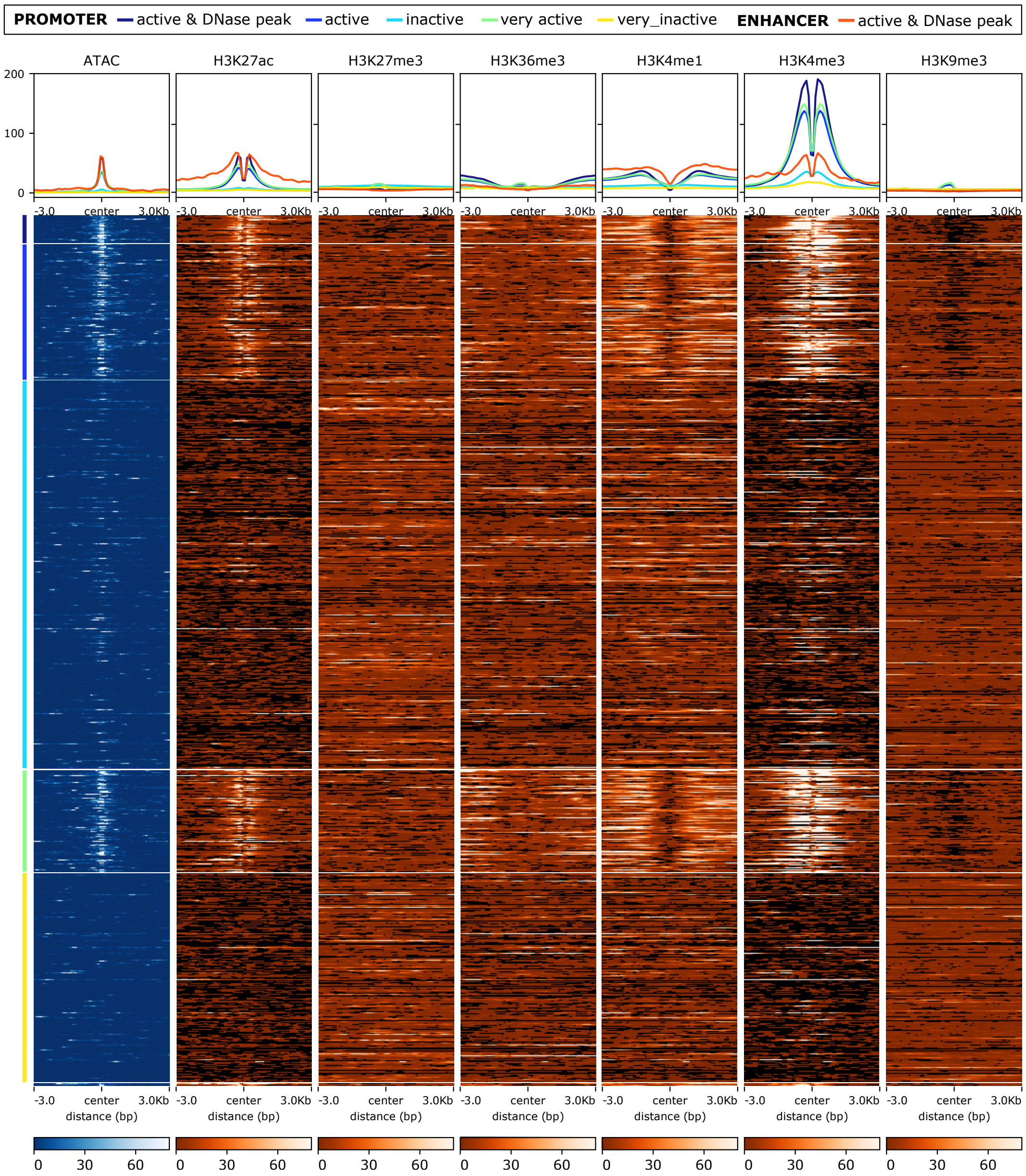


Figure A1: Heatmap of epigenomic data at enhancers and promoters in mESC. Profiles and heatmaps of raw counts in 100 bp bins ±3 kb at active promoters ($\log_2(\text{FPKM}+1) \geq 1$, dark blue), 'very' active promoters ($\log_2(\text{FPKM}+1) \geq 2$, green), active promoters containing a DNase-seq peak ($\log_2(\text{FPKM}+1) \geq 1$, dark blue), inactive promoters ($\log_2(\text{FPKM}+1) < 1$, light blue), 'very' inactive promoters ($\log_2(\text{FPKM}+1) = 0$, yellow), FANTOM5-based active enhancers containing a DNase-seq peak (red). The enhancer set is not shown as heatmap, since it is too small. Plots based on results from *deepTools* (Ramírez *et al.*, 2016).

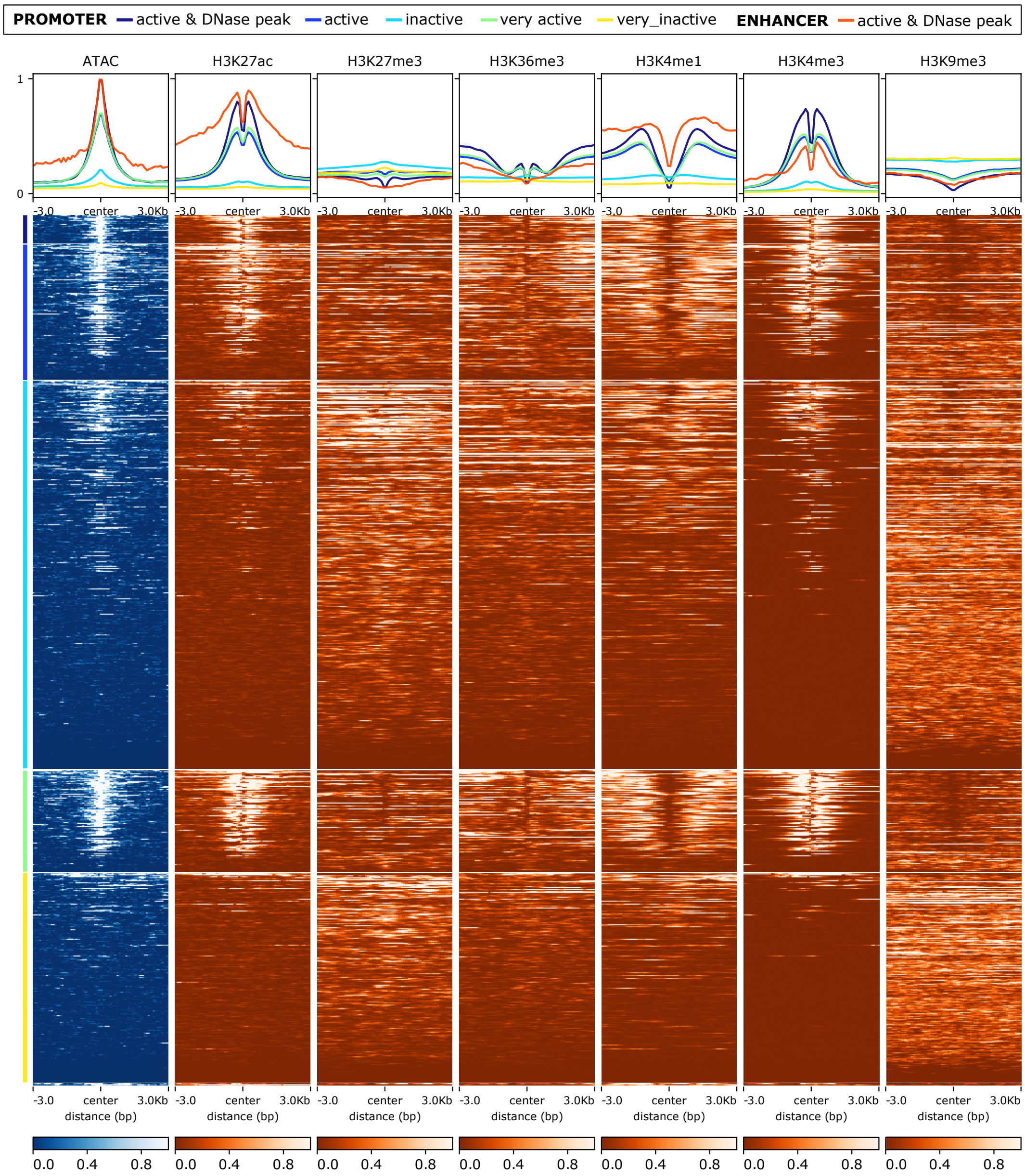


Figure A2: Scaled heatmap of epigenomic data at enhancers and promoters in mESC. Profiles and heatmaps ±3 kb at active promoters ($\log_2(\mathrm{FPKM}{+}1) \geq 1$; dark blue), 'very' active promoters ($\log_2(\mathrm{FPKM}{+}1) \geq 2$; green), active promoters containing a DNase-seq peak ($\log_2(\mathrm{FPKM}+1) \geq 1$, dark blue), inactive promoters ($\log_2(\mathrm{FPKM}+1) < 1$, light blue), 'very' inactive promoters ($\log_2(\mathrm{FPKM}+1) = 0$, yellow), FANTOM5-based active enhancers containing a DNase-seq peak (red). Raw cage counts in 100 bp bins are scaled to $[0, 1]$ for each data individually, after 1st and 95th quantile are excluded. The enhancer heatmap is not shown, since the enhancer set is too small. Plots based on results from *deepTools* (Ramírez *et al.*, 2016).

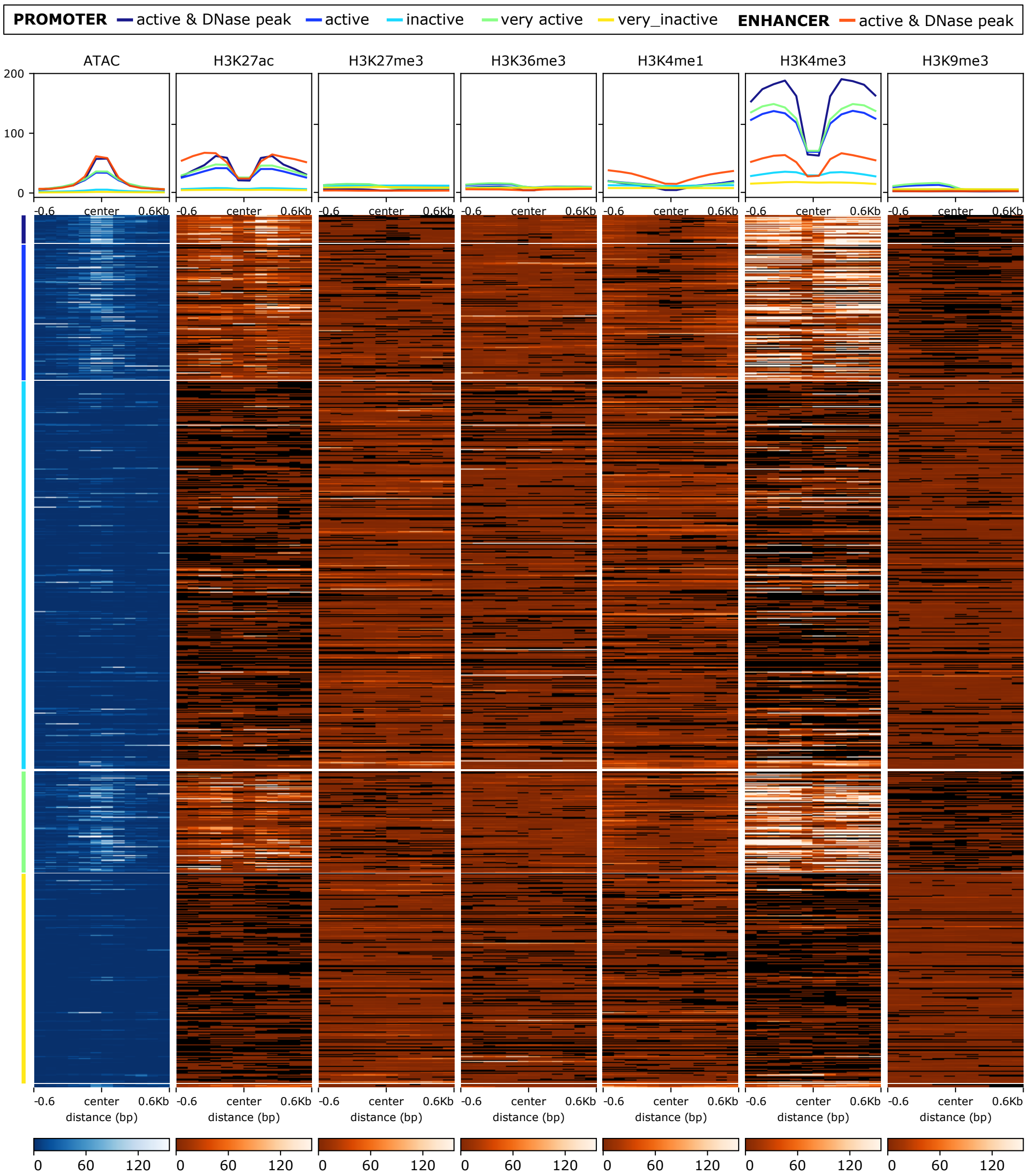


Figure A3: Zoomed heatmap of epigenomic data at enhancers and promoters in mESC. Profiles and heatmaps of raw counts in 100 bp bins ±0.6 kb at active promoters ($\log_2(\text{FPKM} + 1) \geq 1$; dark blue), 'very' active promoters ($\log_2(\text{FPKM} + 1) \geq 2$; green), active promoters overlapping with a DNase-seq peak ($\log_2(\text{FPKM} + 1) \geq 1$, dark blue), inactive promoters ($\log_2(\text{FPKM} + 1) < 1$, light blue), 'very' inactive promoters ($\log_2(\text{FPKM} + 1) = 0$, yellow), FANTOM5-based active enhancers with DNase-seq peak overlap (red). The enhancer set is not shown as heatmap, since it is too small. Plots based on results from *deepTools* (Ramírez *et al.*, 2016).

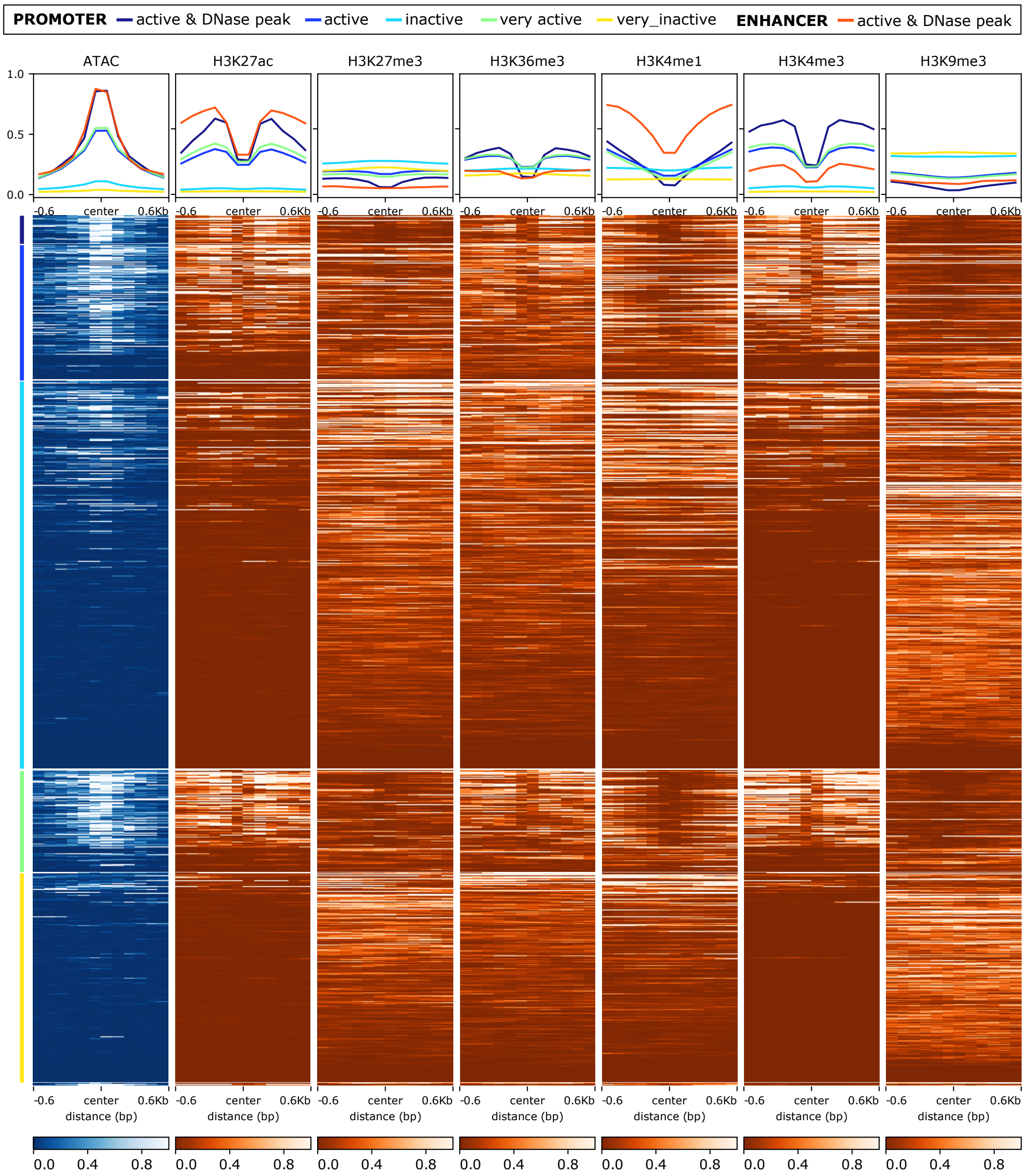


Figure A4: Zoomed and scaled heatmap of epigenomic data at enhancers and promoters in mESC. Profiles and heatmaps ±0.6 kb at active promoters ($\log_2(\mathrm{FPKM}+1) \geq 1$; dark blue), 'very' active promoters ($\log_2(\mathrm{FPKM}+1) \geq 2$; green), active promoters containing a DNase-seq peak ($\log_2(\mathrm{FPKM}+1) \geq 1$, dark blue), inactive promoters ($\log_2(\mathrm{FPKM}+1) < 1$, light blue), 'very' inactive promoters ($\log_2(\mathrm{FPKM}+1) = 0$, yellow), FANTOM5-based active enhancers containing a DNase-seq peak (red). Raw cage counts in 100 bp bins are scaled to $[0, 1]$ for each data individually, after 1st and 95th quantile are excluded. The enhancer heatmap is not shown, since the enhancer set is too small. Plots based on results from *deepTools* (Ramírez *et al.*, 2016).

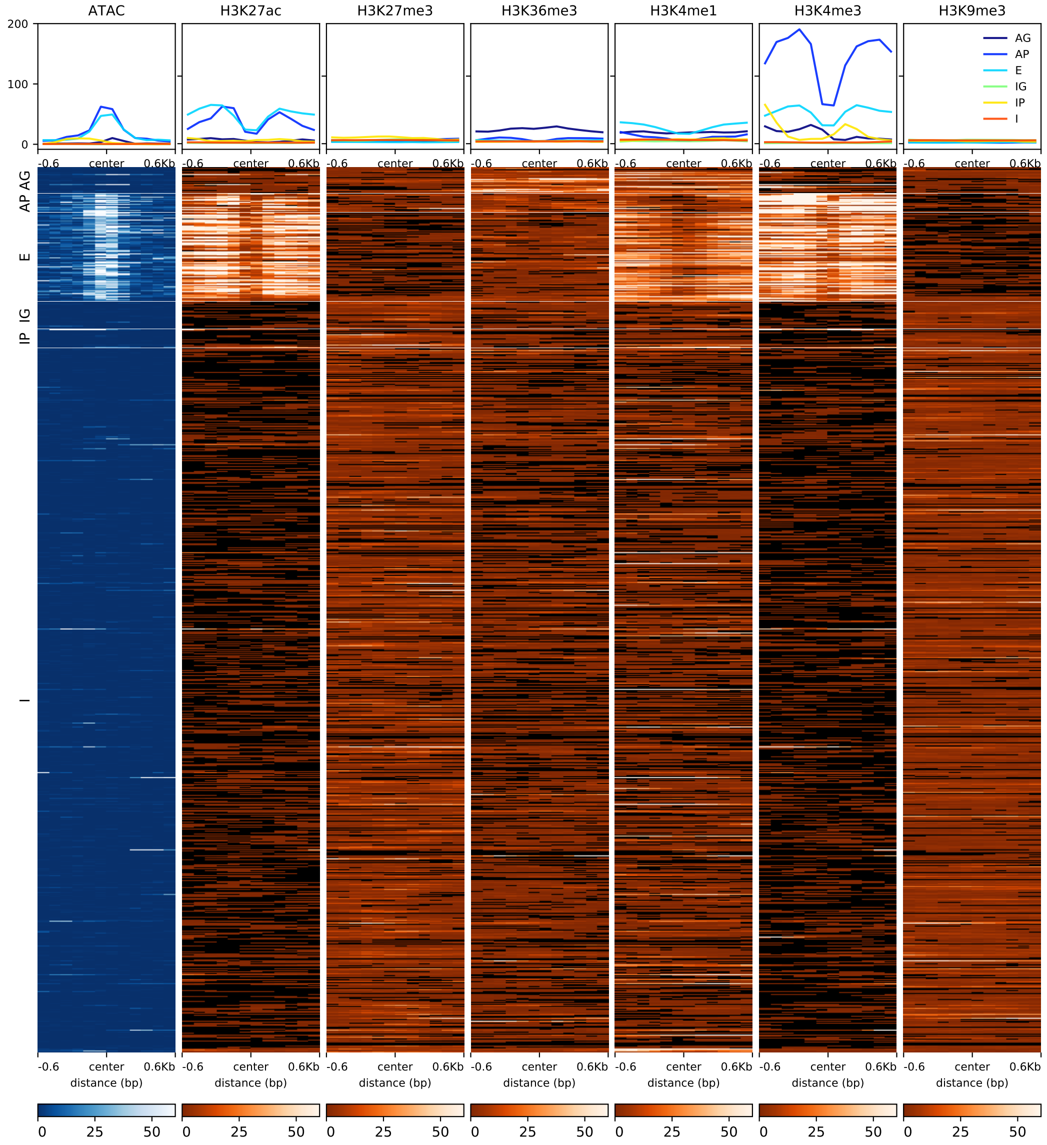


Figure A5: Heatmap of epigenomic data at test set regions. Histone modification and ATAC-seq profiles and heatmaps of raw counts in 100 bp bins for all active genes (AG, dark blue), active promoters (AP, blue), active enhancers (E, light blue), inactive genes (IG, green), inactive promoters (IP, yellow) and intergenic regions (I, red) in the test set. Plots based on results from *deepTools* (Ramírez *et al.*, 2016).

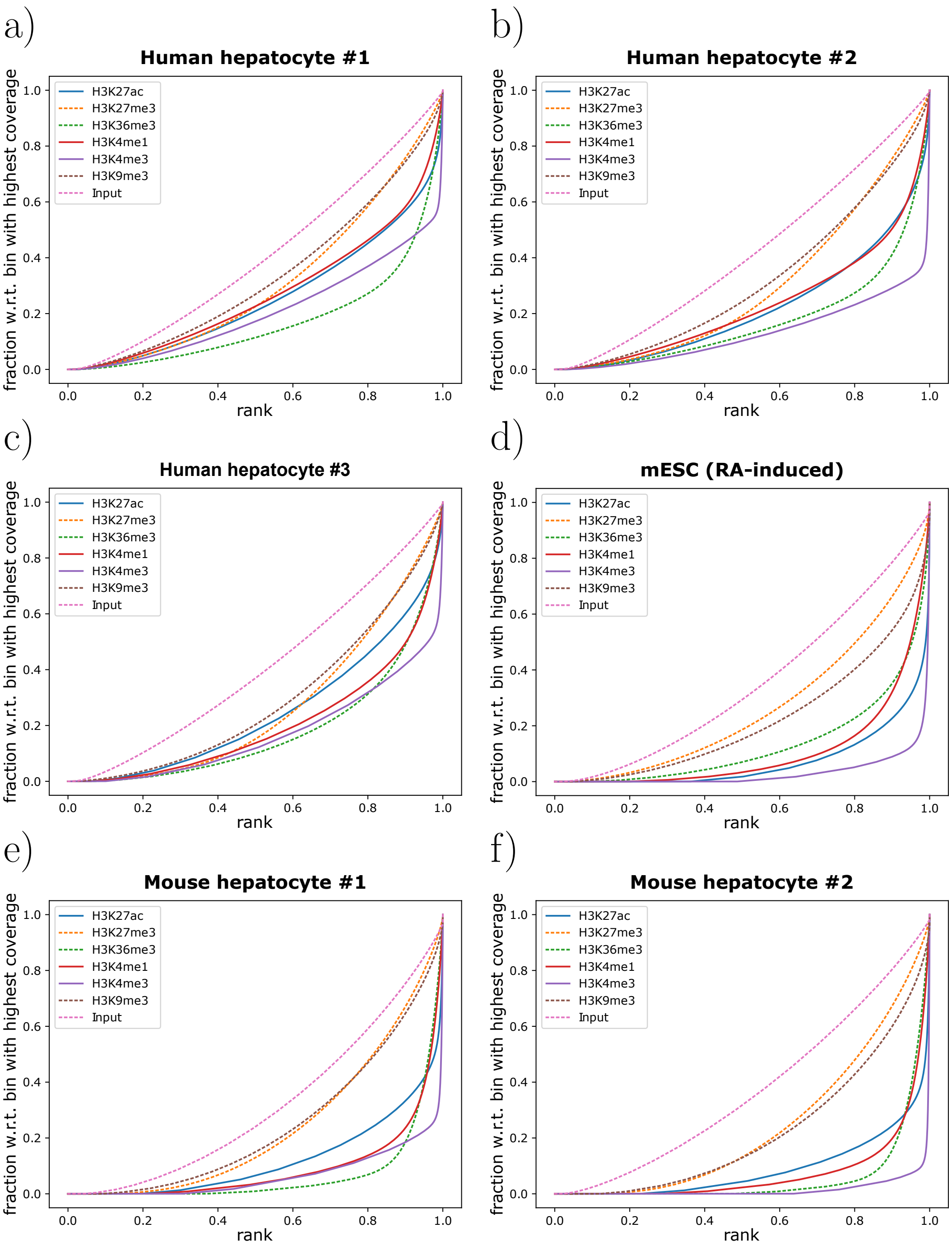


Figure A6: Fingerprint quality control metrics for ChIP-seq experiments. For each HM ChIP-seq data, reads with a mapping quality ≥ 30 are counted per adjacent 500 bp bin. Then, the read counts are sorted and their cumulative sum is plotted. The plots are done with *deepTools* (Ramírez *et al.*, 2016).

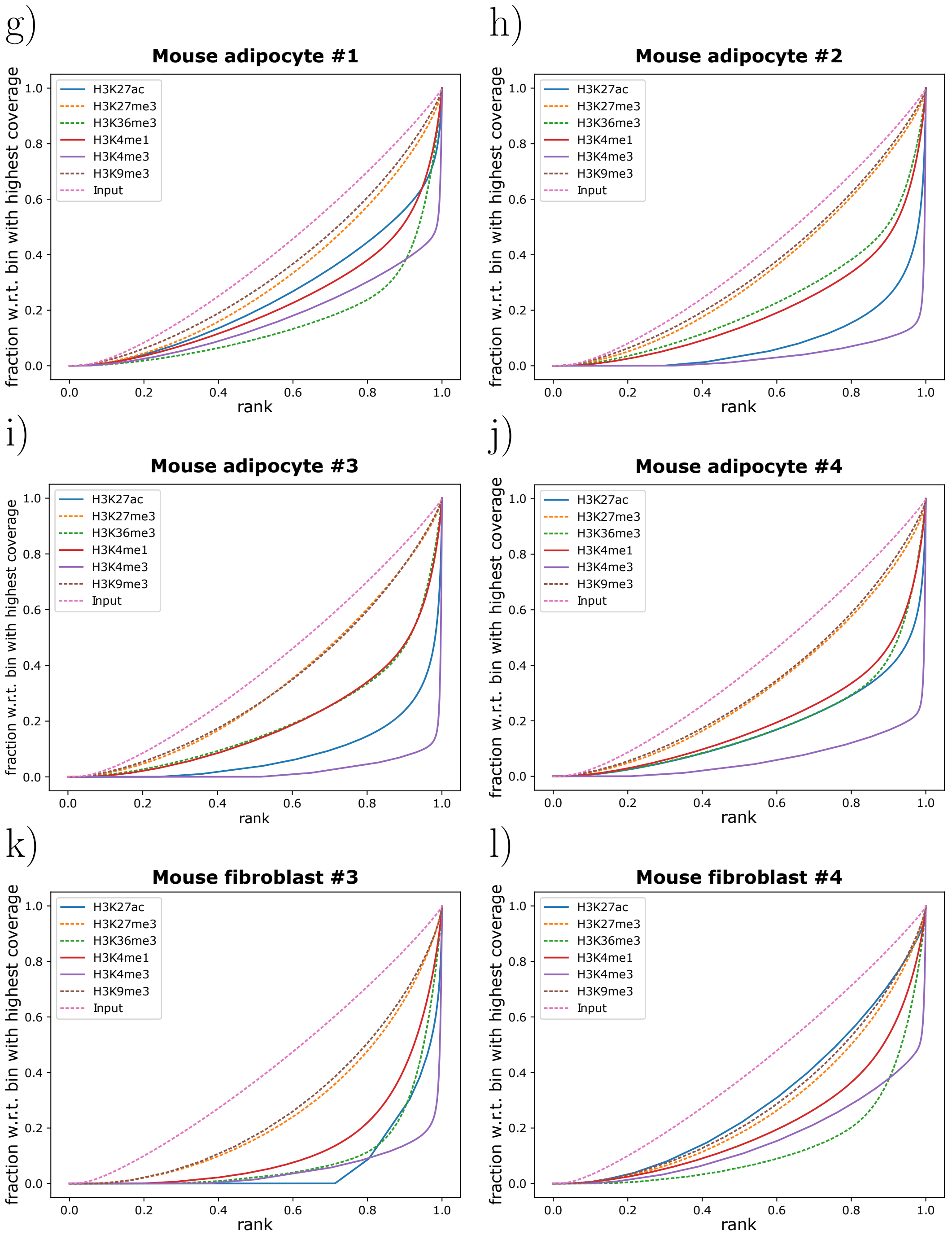
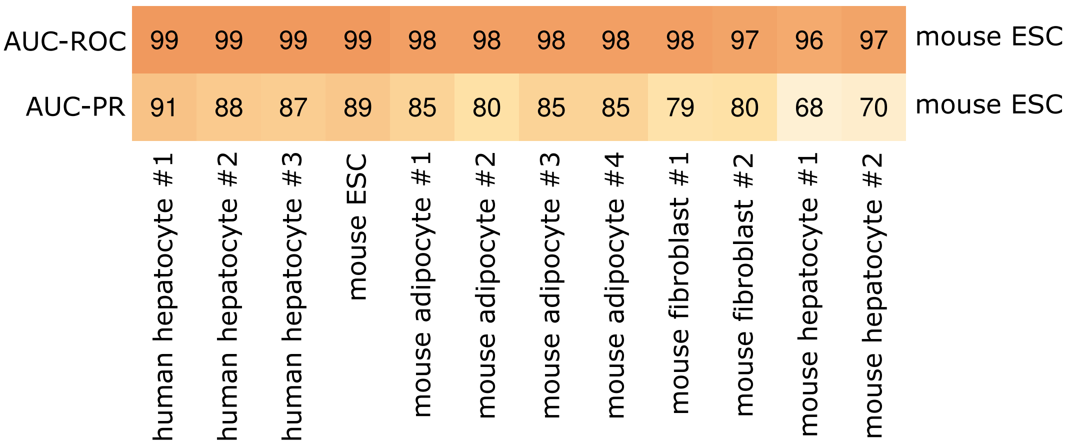


Figure A6: (Cont.) Fingerprint quality control metrics for ChIP-seq experiments.



	human hepatocyte #1	human hepatocyte #2	human hepatocyte #3	mouse ESC	mouse adipocyte #1	mouse adipocyte #2	mouse adipocyte #3	mouse adipocyte #4	mouse fibroblast #1	mouse fibroblast #2	mouse hepatocyte #1	mouse hepatocyte #2	
AUC-ROC	99	99	99	99	98	98	98	98	98	97	96	97	mouse ESC
AUC-PR	91	88	87	89	85	80	85	85	79	80	68	70	mouse ESC

Figure A7: Performance across cell lines and species for REPTILE in original setting. REPTILE was learned using the original p300-based training set and histone modification ChIP-seq data in mouse embryonic stem cells (He *et al.*, 2017). Depicted are AUC-ROC and AUC-PR results for predictions on test sets in 12 samples from different cell lines and species.

Appendix B
Supplementary Tables

Table B1: Enhancer regions defined by FANTOM5. CAGE count data was downloaded for mouse embryonic stem cells, mouse synovial fibroblasts,mouse adipocytes and mouse and human hepatocytes. Depending on the available number of replicates ('$\sum$ *Repl.*') all regions ('# *Regions*') where narrowed down to a set of high confidence enhancers ('*Criterium*'). For example, we used count data of three biological replicates from murine hepatocytes and chose 753 enhancers which had eight and more counts in all three replicates.

Abbr.	FANTOM5 Cell Line Description	Criterium $\left(\sum \text{Repl.}\right)$	# Regions
mESC^{+}	• ES-OS25 embryonic stem cells, DMSO control • ES-OS25 embryonic stem cells, untreated control • ES-Ert2 embryonic stem cells, untreated control , 48hr • ES-OS25 embryonic stem cells, untreated siRNA control • ES-OS25 embryonic stem cells, scrambled siRNA control	≥ 4 counts in all (13)	372
adipocyte	• ST2 (mesenchymal stem cells) cells, differentiation to adipocytes, day06	> 3 counts in all (2)	756
fibroblast (healthy)	• mouse fibroblast cell line: CRL-1658 NIH/3T3	≥ 2 counts in any (1)	683
mouse hepatocyte	• liver sinusoidal endothelial cells, partial hepatectomy, 01week	≥ 8 counts in all (3)	753
human hepatocyte	• liver, adult, pool1	≥ 8 counts in any (1)	298

Table B2: Final enhancer regions defined from FANTOM5 and DNaseI peaks. DNase-seq peaks were called for each sample/replicate. The overlap of DNAseI peaks and the filtered FANTOM5 regions from Table B1 build the final enhancer lists used in our workflow. Here, for each type of tissue, we chose the overlap set with the maximal size (bold). For example, we used the 239 DNaseI peaks of sample 1 that overlap with FANTOM5 as representative enhancer set for mouse hepatocytes.

Abbreviation	sample/replicate	# DNaseI peaks	# FANTOM5	# overlap
mESC$^+$	replicate 1	123576	372	**280**
	replicate 2	88973		250
adipocyte	sample 1	43814	756	**292**
fibroblast (healthy)	sample 1	90858	683	**251**
	sample 2	65682		141
mouse hepatocyte	sample 1	51110	753	**239**
	sample 2	44336		227
human hepatocyte	sample 1	86296	2472	**339**
	sample 2	44290		224
	sample 3	40438		234

Table B3: Active promoter regions defined by RNA-seq cutoff and DNase peaks. We expanded the TSSs of active genes (”active” according to definition in Section 4.1.6) symmetrically to a total length of 100 bp and computed the overlap with DNAse summits in the same tissue (not always the same sample). Only expanded TSSs containing a DNase summit are finally used to define active promoters.

sample/replicate	# active promoter	DNaseI sample	# overlap
mESC$^+$	10,044	replicate 1	2,853
adipocyte sample 1	9,217	sample 1	2,273
adipocyte sample 2	9,206	sample 1	2,295
adipocyte sample 3	9,245	sample 1	2,317
adipocyte sample 4	9,317	sample 1	2,339
fibroblast (healthy) sample 1	10,593	sample 1	2650
fibroblast (healthy) sample 2	10,326	sample 1	2528
mouse hepatocyte sample 1	8,392	sample 1	2,299
mouse hepatocyte sample 2	8,417	sample 1	2,318
human hepatocyte sample 1	6,668	sample 1	1,689
human hepatocyte sample 2	6,853	sample 1	1,749
human hepatocyte sample 3	7,021	sample 1	1,670

Table B4: JASPAR CORE vertebrates clustering. Composition of all JASPAR clusters used in this work which consist of more than one motif.

cluster	motif name
1	DUXA, DUX4, PROP1, Phox2b, PHOX2A, Arid3a, HOXA5, Crx, RHOXF1, OTX1, Pitx1, OTX2, PITX3, GSC, GSC2, Nkx2-5, NKX2-3, NKX2-8, ISL2, NKX3-2, Nkx3-1, HMBOX1, Arid3b, Lhx3, Dux, BARHL2, Barhl1, POU6F2, VENTX, Nobox, LBX1, PDX1, NKX6-1, NKX6-2, BARX1, BSX, EN1, LHX9, ISX, Shox2, SHOX, RAX2, Prrx2, PRRX1, UNCX, Dlx2, DLX6, Dlx3, Dlx4, MSX2, MSX1, Msx3, PAX4, Lhx8, LMX1B, LMX1A, Lhx4, VAX1, VAX2, VSX1, VSX2, mix-a, POU6F1, LHX6, LHX2, NOTO, MNX1, Dlx1, EN2, ALX3, MIXL1, GBX1, GBX2, RAX, HESX1, ESX1, LBX2, MEOX1, MEOX2, GSX1, GSX2, HOXA2, HOXB2, HOXB3, EVX1, EVX2, EMX1, EMX2
6	TFAP2B_var.2, TFAP2C_var.2, TFAP2B_var.3, TFAP2C_var.3, TFAP2A_var.3, TFAP2A_var.2, TFAP2C, TFAP2B
10	ONECUT3, ONECUT1, ONECUT2, PAX7, PAX3, CUX1, CUX2
12	FOXP3, FOXI1, FOXO4, FOXO6, FOXL1, FOXD2, FOXO3, FOXD1, FOXG1, Foxj2, SRY, Sox5, Foxq1, Foxd3, FOXF2, Foxj3, FOXA1, FOXK1, Foxo1, FOXP1, FOXP2, FOXK2, Foxa2, FOXC2, FOXC1, FOXB1
18	LIN54, Pou5f1::Sox2, POU3F4, POU5F1B, Pou2f3, POU2F1, POU1F1, POU3F3, POU3F1, POU3F2, POU2F2, POU5F1
20	MEF2C, MEF2B, MEF2A, MEF2D
22	PBX1, Hoxc9, Hoxa9, HOXC10, HOXD11, HOXC13, Hoxa11, HOXC11, HOXD12, HOXC12, CDX2, CDX1, HOXA10, Hoxd9, HOXD13, HOXA13, HOXB13
24	ZIC1, ZIC3, ZIC4, GLIS1, GLIS2, GLIS3, INSM1, ZNF740, MZF1_var.2
28	SP2, KLF13, SP4, KLF14, Klf12, SP8, KLF16, SP3, KLF5, SP1, KLF4, Klf1, KLF9, GLI2, ZBTB7B, ZBTB7C, EGR2, EGR4, EGR1, EGR3
30	HNF1B, HNF1A, POU4F1, POU4F2, POU4F3
33	SOX13, Sox17, SOX10, SOX15, Sox3, Sox6, Sox2, SOX9
35	NFIC::TLX1, NFIC, NFIA, NFIX, THAP1, HIC2, Hic1
41	PPARG, ESR2, ESR1
48	CTCFL, CTCF
54	EWSR1-FLI1, ZNF263

Appendix C
Data Processing

For all experimental data used in this work, ChIP was performed against H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K36me3 and H3K9me3, and the sheared chromatin without antibody (Input) served as a control. Moreover, duplicates were removed using Picard tools (Wysoker *et al.*, 2013) for all ChIP-seq experiments. All DNase-seq experiments were aligned with BWA-MEM (Li *et al.*, 2013) to the genome assembly as indicated in Table C1, and duplicates were removed using Picard tools (Wysoker *et al.*, 2013). More detailed information, for example on cell culture, isolation or sequencing libraries, is described below. The used alignment methods are summarized in Table C1.

Table C1: Overview of alignment methods. Raw sequencing reads were aligned to the indicated genome assembly ('GRCm38' or 'hs37d5') with the indicated alignment method (BWA-MEM (Li *et al.*, 2013), STAR (Dobin *et al.*, 2012) or TopHat2 (Kim *et al.*, 2013).

Tissue/Cell type	genome assembly	ChIP-seq	RNA-seq
Mouse embryonic stem cells	'GRCm38'	STAR	BWA-MEM
Mouse synovial fibroblasts	'GRCm38'	BWA-MEM	TopHat2
Mouse adipocytes	'GRCm38'	BWA-MEM	TopHat2
Mouse hepatocytes	'GRCm38'	STAR	BWA-MEM
Human hepatocytes	'hs37d5'	BWA-MEM	TopHat2

Mouse embryonic stem cells

E14 mouse embryonic stem cells (mESCs) were cultured and routinely passaged every two days in ES medium plus leukemia inhibitory factor (LIF) in order to maintain the pluripotent state of the cells (Smith *et al.*, 1988; Pease *et al.*, 1990)). To exit from pluripotency and push the cells towards differentiation, LIF was withdrawn and retinoic acid (RA) was added to the medium for a short pulse of 4h. For ChIP-seq, cells were harvested, fixed and processed according to the standard ChIP protocol (Ramisch *et al.*, 2018). Sequencing libraries were prepared and the resulting DNA fragments were paired-end 50bp sequenced on a Illumina HiSeq 2500 device. For RNA-seq, cells were harvested and three biological replicates were subjected to RNA extraction. Sequencing libraries were generated from total mRNA input and high-throughput sequencing was performed on an Illumina HiSeq 2500 device generating resulting in 50bp paired-end reads. For ATAC-seq, cells were subjected to transposition reaction and PCR amplification of accessible regions by Omni-ATAC-seq (Corces *et al.*, 2017). Sequencing libraries were constructed and DNA fragments were paired-end 50bp sequenced on a Illumina *HiSeq 4000* device. Raw reads were aligned to the genome assembly *'GRCm38'* using BWA-MEM (Li *et al.*, 2013) and duplicates were removed with SAMtools (Li *et al.*, 2009). ATAC-seq peaks were identified using MACS2 (Zhang *et al.*, 2008). Raw reads from DNase-seq experiments were downloaded from GEO (GSM1014154). All experimental ChIP-seq, RNA-seq and ATAC-seq data related to these samples are accessible on GEO (accession Nr.: GSE120376).

Mouse synovial fibroblasts

Mouse synoial fibroblasts were isolated by enzymatic digestion from hind limbs of 12 week old *hTNFtg* (reactive arthritis, strain Tg197 overexpressing human TNF) and wildtype (healthy control) (Wehmeyer *et al.*, 2016; Keffer *et al.*, 1991). ChIP-seq was carried out as described in Arrigoni *et al.* (2016). Resulting DNA fragments were paired-end 50bp sequenced on a Illumina HiSeq 2500 device. For RNA-seq, long RNA libraries were prepared from total mRNA input and sequenced on an Illumina HiSeq 2500 device resulting in 50bp and

100bp long paired-end reads. For DNase-seq, nuclei were digested with DNaseI in five different dilutions as described by Schmidt (2016).

Mouse adipocytes

Samples for adipocytes were isolated by collagenase treatment for five minutes followed by five minutes of collagenase inactivation (Arrigoni *et al.*, 2016). After centrifugation the fat layer was collected. For ChIP-seq, chromatin from fixed cells has been extracted and sonicated for 15 minutes using Covaris S220 sonicator. Resulting DNA fragments were paired-end 50 bp sequenced on an Illumina HiSeq HiSeq 2500 device. For RNA-seq, RNA isolation for cells was performed using 1 ml TRIzol per sample followed by Isopropyl alcohol/Ethanol precipitation. Sequencing libraries were generated from total mRNA input and sequenced on an Illumina HiSeq 2500 device resulting in 100bp paired-end reads. For DNase-seq, nuclei extracted from $\sim 10 \times 10^6$ cells by treatment with IGEPAL were digested with different concentrations of DNaseI (Schmidt, 2016). Sequencing libraries were prepared and sequenced on an Illumina HiSeq 2500 device resulting in 100bp long paired-end reads.

Mouse hepatocytes

Primary mouse hepatocytes were obtained from two female mice (C57BL/6J x DBA/2 background) at the age of nine weeks. The isolation of primary mouse hepatocytes was performed by a two-step EDTA/collagenase perfusion technique (Godoy *et al.*, 2013). ChIP-seq was performed using primary mouse hepatocytes as described in Kinkley *et al.* (2016) with minor modifications, and libraries from each sample were pooled and paired-end sequenced on an HiSeq 2500 device. For RNA-seq, RNA isolation for cells was performed using 1 ml TRIzol per sample followed by Isopropyl alcohol/RNA was extracted from hepatocytes homogenized in 1 mL Trizol. Sequencing libraries were generated from total mRNA input using TruSeq v3 Kit (Illumina) according to manufacturer's instructions and seqeunced on an Illumina HiSeq 2500 device resulting in 100bp paired-end reads. For DNase-seq, nuclei extracted from $\sim 10 \times 10^6$

cells by treatment with IGEPAL were digested with different concentrations of DNaseI (Schmidt, 2016). Sequencing libraries were prepared and sequenced on an Illumina HiSeq 2500 device resulting in 100bp long paired-end reads.

Human hepatocytes

Primary human hepatocytes were obtained from three different female donors (age 28-70 years) undergoing surgery due to primary or secondary liver tumors. Hepatocytes were isolated from healthy liver tissue remaining from liver resection (Godoy *et al.*, 2013). ChIP-seq was performed using primary human hepatocytes as described in Kinkley *et al.* (2016) with minor modifications, and libraries from each sample were pooled and paired-end sequenced on an HiSeq 2500 device. For RNA-seq, RNA was extracted from hepatocytes homogenized in 1 mL Trizol. Sequencing libraries were generated from total mRNA input using TruSeq v3 Kit (Illumina) according to manufacturer's instructions and seqeunced on an Illumina HiSeq 2500 device resulting in 100bp paired-end reads. For DNase-seq, nuclei extracted from $\sim 10 \times 10^6$ cells by treatment with IGEPAL were digested with different concentrations of DNaseI (Schmidt, 2016). Sequencing libraries were prepared and sequenced on an Illumina HiSeq 2500 device resulting in 100bp long paired-end reads.

Processing of HiC-seq experiments

The Juicertools command '*dump*' (Durand *et al.*, 2016) was used to extract data from Hi-C archives associated with mESCs (Bonev *et al.*, 2017):
`http://hicfiles.s3.amazonaws.com/external/bonev/ES_mapq30.hic`
The Hi-C data matrix is Knight-Ruiz normalized (Knight and Ruiz, 2013) at $10kb$ resolution. Topologically associated domains were identified by applying TopDom (Shin *et al.*, 2016) on the $25kb$ binned and normalized matrix with a window size of 750 kb (30 × 25kb). The resulting regions were used to reduce the search space for promoter/gene-enhancer interactions.

Appendix D
Zusammenfassung

In dieser Doktorarbeit zeigen wir, wie man die aktuellen Enhancer-Kentnisse nutzen und verschiedene epigenetische Datensätze integrieren kann um die Postition aktiver Enhancer unter spezifischen Bedingungen vorherzusagen.

Zuerst stellen wir eine neue Methode zur genomweiten Enhancer-Vorhersage basierend auf Histonmodifikationsdaten vor. Unsere Methode kombiniert zwei Random Forest Klassifikationsverfahren zur Unterscheidung von aktiven und inaktiven genomischen Regionen und zur schwierigeren Unterscheidung von aktiven Enhancern und aktiven Promotoren. Beim Modellieren und Optimieren der Klassifikationsmerkmale (Feature) berücksichtigen wir die lokale Chromatinstruktur. Kennzeichnend für einen aktiven Enhancer ist im Wesentlichen ein Abschnitt zugänglichen Chromatins, umgeben von Nukleosomen mit spezifischen Histonmodifikationen. Unsere Trainings-Enhancer sind so definiert, dass sie offene Chromatinregionen umfassen und nachweislich bidirektionale Transkripte herstellen. Diese Enhancer-Charakteristiken haben wir möglichst unabhängig von den Klassifikationsmerkmalen gewählt um Zirkelschlüsse zu vermeiden. Wir haben unsere Methode in embryonalen Stammzellen der Maus validiert und sehr gute Vorhersagergebnisse auf ausgewählten Testsets erzielt. Außerdem haben wir vorhergesagte, beieinanderliegende Enhancer in Regionen hoher Enhancer-Dichte zusammengefasst, für die wir eine gute Übereinstimmung mit veröffentlichten Superenhancern feststellen konnten. Im Gegensatz zu vielen Methoden zur Enhancer-Vorhersage bieten wir ein trainiertes Modell mit integriereter Datennormalisierung an, dass zuverlässig auf neue Datensätze anderer Zelltypen und Spezies angewendet werden kann. Unser Modell zeigt bessere Ergenisse als die viel genutzte Methode ChromHMM, und ist bei Anwendung innerhalb eines Zelltyps vergleichbar mit der REPTILE-Methode. Für die Anwendung auf neue Datensätze ist unsere Methode besser geeignet.

Schließlich zeigen wir, wie unser trainiertes Modell als Basis eines Frameworks fungieren kann um bedingungsspezifische regulatorische Einheiten (Enhancer-Gen-Paare) von Histonmodifikations- und Genexpressionsdaten vorherzusagen.

Appendix E
Summary

In this thesis, we show how to exploit the current knowledge of enhancers, and integrate different types of epigenomic data to make condition-specific predictions on the location of active enhancers.

First, we introduce a novel method for genome-wide enhancer prediction which is solely based on histone modification data. Our method is a combination of two random forest classifiers, where one classifier learns the difference between active and inactive genomic regions and the other concentrates on the more difficult task to distinguish active enhancers from active promoters. We model and optimize the corresponding features taking into account the local chromatin structure. For an active enhancer, this is in essence an accessible region flanked by nucleosomes with specific histone modifications. To avoid circular reasoning, our training enhancers are defined by feature set-independent characteristics: accessibility and bidirectional transcription. We thoroughly validate our method on mouse embryonic stem cell data and achieve very good performances on a constructed test set as well as on a validated set of enhancers. Moreover, our genome-wide enhancer predictions have a high spatial resolution. We also cluster proximal enhancers and show that the resulting regions of high enhancer density are in good agreement with a published list of super-enhancers in mouse embryonic stem cells. In contrast to many other methods, we offer a pre-trained classifier with integrated data normalization that can be used to reliably predict enhancers across different cell types and species. This classifier is superior to the prominent unsupervised method ChromHMM, and shows similar results as the recent supervised REPTILE approach when applied in the same cell type. In terms of transferability to other conditions, our method outperforms REPTILE.

Finally, we demonstrate how our pre-trained classifier can be embedded into a comprehensive framework to predict condition-specific regulatory units (pairs of enhancers and putative target genes) of histone modification and gene expression data.

Appendix F
Short Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

For reasons of data protection, the curriculum vitae is not included in the online version.

Publications

Ramisch, A.*, Heinrich, V.*, Glaser, L. V., Fuchs, A., Yang, X., Benner, P., Schoepflin, R., Li, N., Kinkley, S., Hillmann, A., Longinotto, J., Heyne, S., Czepukojc, B., Kessler, S. M., Kiemer, A. K., Cadenas, C., Arrigoni, L., Gasparoni, N., Manke, T., Pap, T., Pospisilik, A., Hengstler, J., Walter, J., Meijsing, S. H., Chung, H.-R., and Vingron, M. (2018). CRUP: A comprehensive framework to predict condition-specific regulatory units. *bioRxiv* , page 501601. (*equal contribution)

Grasse, S., Lienhard, M., Frese, S., Kerick, M., Steinbach, A., Grimm, C., Hussong, M., Rolff, J., Becker, M., Dreher, F., Schirmer, U., Boerno, S., **Ramisch, A.**, Leschber, G., Timmermann, B., Grohe, C., Luders, H., Vingron, M., Fichtner, I., Klein, S., Odenthal, M., Buttner, R., Lehrach, H., Sultmann, H., Herwig, R., and Schweiger, M. R. (2018). Epigenomic profiling of non-small cell lung cancer xenografts uncover LRP12 DNA methylation as predictive biomarker for carboplatin resistance. *Genome Med.*, **10**(1), 55.

Sheinman, M., **Ramisch, A.**, Massip, F., and Arndt, P. F. (2016). Evolutionary dynamics of selfish DNA explains the abundance distribution of genomic subsequences. *Sci. Rep.*, **6**, 30851.

Huska, M. R.*, **Ramisch, A.***, Vingron, M., and Marsico, A. (2016). Predicting enhancers using a small subset of high confidence examples and co-training. *PeerJ Preprints*, **4**, e2407v1. (*equal contribution)