Prediction of transcription factor binding using epigenomics and genome variation data

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 12
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

Nina Baumgarten
aus Kaiserslautern

Frankfurt am Main, 2023

Vom Fachbereich 12 der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:

Gutachter:

Datum der Disputation:

Abstract

A central concern in genetics is to identify mechanisms of transcriptional regulation. The aim is to unravel the mapping between the DNA sequence and gene expression. However, it turned out that this is extremely complex. Gene regulation is highly cell type-specific and even moderate changes in gene expression can have functional consequences.
Important contributors to gene regulation are transcription factors (TFs), that are able to directly interact with the DNA. Often, a first step in understanding the effect of a TF on the gene's regulation is to identify the genomic regions a TF binds to. Therefore, one needs to be aware of the TF's binding preferences, which are commonly summarized in TF binding motifs. Although for many TFs the binding motif is experimentally validated, there is still a large number of TFs where no binding motif is known. There exist many tools that link TF binding motifs to TFs. We developed the method Massif that improves the performance of such tools by incorporating a domain score that uses the DNA binding domain of the studied TF as additional information.
TF binding sites are often enriched in regulatory elements (REMs) such as promoters or enhancers, where the latter can be located megabases away from its target gene. However, to understand the regulation of a gene it is crucial to know where the REMs of a gene are located. We introduced the EpiRegio webserver that holds REMs associated to target genes predicted across many cell types and tissues using *STITCHIT*, a previously established method. Our publicly available webserver enables to query for REMs associated to genes (gene query) and REMs overlapping genomic regions (region query). We illustrated the usefulness of EpiRegio by pointing to a TF that occurs enriched in the REMs of differential expressed genes in circPLOD2 depleted pericytes. Further, we highlighted genes, which are affected by CRISPR-Cas induced mutations in non-coding genomic regions using EpiRegio's region query.
Non-coding genetic variants within REMs may alter gene expression by modifying TF binding sites, which can lead to various kinds of traits or diseases. To understand the underlying molecular mechanisms, one aims to evaluate the effect of such genetic variations on TF binding sites. We developed an accurate and fast statistical approach, that can assess whether a single nucleotide polymorphism (SNP) is regulatory. Further, we combined this approach with epigenetic data and additional analyses in our Sneep workflow. For instance, it enables to identify TFs whose binding preferences are affected by the analyzed SNPs, which is illustrated on eQTL datasets for different cell types. Additionally, we used our Sneep workflow to highlight cardiovascular disease genes using regulatory SNPs and REM-gene interactions.
Overall, the described results allow a better understanding of REM-gene interactions and their interplay with TFs on gene regulation.

Publications

In the process of the research that lead to this thesis, the following papers were published. If a work is still in the process of getting published, we provided it as preprint in the list below. Parts of the publications and preprints are included within this thesis.

* shared first authorship

- **N. Baumgarten**, L. Rumpf, T. Kessler and M.H. Schulz; **A statistical approach to identify regulatory DNA variations**, bioRxiv, February 2023, https://doi.org/10.1101/2023.01.31.526404

- C. Zhu*, **N. Baumgarten***, M.Wu, Y. Wang, A.P. Das, J. Kaur, F.B. Ardakani, T.T Duong, M.D. Pham, M. Duda, S. Dimmeler, T, Yuan, M.H. Schulz and J. Krishnan; **CVD-associated SNPs with regulatory potential drive pathologic non-coding RNA expression** bioRxiv, February 2023, https://doi.org/10.1101/2023.02.12.528184

- **N. Baumgarten***, F. Schmidt*, M. Wegner, M. Hebel, M. Kaulich and M.H. Schulz; **Computational prediction of CRISPR-impaired non-coding regulatory regions**, Biological Chemistry, July 2021, https://doi.org/10.1515/hsz-2020-0392

- **N. Baumgarten***, D. Hecker*, S. Karunanithi*, F. Schmidt, M. List and M.H. Schulz; **EpiRegio: analysis and retrieval of regulatory elements linked to genes**, Nucleic Acids Research, July 2020, https://doi.org/10.1093/nar/gkaa382

- **N. Baumgarten**, F.Schmidt and M.H. Schulz; **Improved linking of motifs to their TFs using domain information**, Bioinformatics, March 2020, https://doi.org/10.1093/bioinformatics/btz855

- F. Schmidt, A. Marx, **N. Baumgarten**, M. Hebel, M. Wegner, M. Kaulich, M.S. Leisegang, R.P. Brandes, J. Göke, J. Vreeken and M.H. Schulz; **Integrative analysis of epigenetics data identifies gene-specific regulatory elements**, Nucleic Acids Research, October 2021, https://doi.org/10.1093/nar/gkab798

- S.F. Glaser, A.Brezski, **N. Baumgarten**, M. Klangwart, A.W. Heumüller, R.K Maji, M.S. Leisegang, S. Guenther, C.M. Zehendner, D. John, M.H. Schulz, K. Zarnack and S. Dimmeler; **Circular RNA circPLOD2 regulates pericyte function by targeting the transcription factor**

KLF4, bioRxiv, December 2022,
https://doi.org/10.1101/2022.12.04.519017

- P. Cattaneo, M.G.B. Hayes, **N. Baumgarten**, D. Hecker, *et al*; **DOT1L regulates chamber-specific transcriptional networks during cardiogenesis and mediates postnatal cell cycle withdrawal**, Nature Communications, December 2022,
https://www.nature.com/articles/s41467-022-35070-2

- L.S. Tombor, D. John, S.F. Glaser, G. Luxán, E. Forte, M. Furtado, N. Rosenthal, **N. Baumgarten**, M.H. Schulz, *et al*; **Single cell sequencing reveals endothelial plasticity with transient mesenchymal activation after myocardial infarction**, Nature Communications, 12(1): 681. January 2021, https://doi.org/10.1038/s41467-021-20905-1

Table of Contents

Chapter 1

Introduction

Molecular genetics aims to answer the question of how an organism decodes the genetic and epigenetic information into a phenotype. Starting from a single zygote, the human body is evolved consisting of trillions of cells, which can be classified into roughly 200 different cell types (Roy and Conroy, 2018). Each cell type fulfills specific tasks. For instance, cardiomyocytes are responsible for the contraction of the hearth (Keepers *et al.*, 2020). Hepatocytes, the most common cell type in the liver, are involved, among other functionalities, in the detoxification of the blood (Schulze *et al.*, 2019). Further examples are T lymphocytes, which are known to play a role in the immune response of the body by establishing cell immunity (Campbell and Reece, 2009, p. 1290) or rods and cones, cells of the eye, which contribute to the human vision (Lamb, 2015).

How do cells establish their different functionalities, when they all rely on the same DNA? The expression of genes differs from cell to cell, not only in how strong a gene is transcribed, but also in which genes are expressed. Gene expression is regulated by transcription factors (TFs), proteins that bind to the DNA by recognizing specific DNA patterns, known as motifs (Lambert *et al.*, 2018). These motifs occur enriched in regulatory elements (REMs) such as promoters or enhancers (Gasperini *et al.*, 2020). Active REMs, which are bound by TFs and their co-factors, influence gene expression. Promoters, which are genomic regions upstream of the gene's transcription start site, initiate the transcription. In contrast, enhancers support the expression by communicating with the promoter via 3D loops even if they are thousands of base pairs away.

With the emergence of high throughput next generation sequencing (NGS) and genomics technologies, the hope was to decipher the so-called cis-regulatory

code, which describes the mapping from the DNA sequence to gene expression (Zeitlinger, 2020; Kim and Wysocka, 2023). However, it turned out that this is not as easy as for the genetic code, which enables to predict the amino acid sequence from the DNA sequence. One reason is that REMs are highly cell type-specific, where some of them are only active at defined time points such as in embryonic development (Levine and Davidson, 2005). Further, it is not clear where in the DNA sequence the REMs that regulate a gene are located. Some of the REMs are known to be megabases away from the promoter of the gene they interact with (Long *et al.*, 2020; Lettice, 2003) or even interact across chromosomes (Monahan *et al.*, 2019). Additionally, the quantitatively precise regulation of genes is extremely important, because even moderate changes in the strength of the expression can lead to various kinds of diseases such as cancer (Flavahan *et al.*, 2015; Cho *et al.*, 2018) or cardiovascular diseases (reviewed in (Yoshida *et al.*, 2019)). Thus, a central concern of genetics is to understand how genes are regulated. Throughout this thesis we aim to contribute by helping to answer the following questions:

- How to precisely predict the motif that describes a TFs' binding behavior? Is it possible to improve existing methods by incorporating the DNA binding domain? (Chapter 3)
- How to detect REMs and their associated target genes? How to identify those REMs active in the studied cell type or tissue? Is such information publicly available and how to easily access it? (Chapter 4)
- How to assess non-coding genetic variations? What is their impact on TF binding sites and how do they affect gene expression? (Chapters 5 and 6)

The content outlined in this thesis was developed within projects either of our group or in cooperation with other labs. Thus, for each presented paper we provide a detailed list with the authors' contribution. The methods developed throughout this thesis are written in UPPERCASE letters (MASSIF, EPIREGIO, and SNEEP).

Chapter 2

Background

This chapter provides background knowledge which might be helpful to understand the methods developed and analyses performed throughout the thesis. It is separated into two parts: In Section 2.1 an overview of biological concepts and techniques is given. The second part of this chapter introduces the computational and mathematical concepts used within this thesis.

2.1 Biological basics

In this section we shortly introduce DNA, RNA, proteins and outline a current definition of genes and how to measure their expression. Further, we provide an overview about chromatin structure and accessibility, epigenetic modifications and gene regulation. The typically used experiments to analyze open chromatin, and to detect regions occupied by histones or DNA-binding proteins are explained as well. Next, we outline a few resources that provide publicly available functional genomics data resulting from international effort. Finally, we summarize the CRISPR-Cas technology.

2.1.1 DNA and RNA

The carrier of the genetic information is the deoxyribonucleic acid (DNA), a macro molecule that is composed of thousands of nucleotides. Each nucleotide consists of three components: (1) a purine- or pyrimidine base, (2) a five-carbon sugar, which in case of DNA is the deoxyribose and (3) a phosphate group. To build a nucleotide, the C1 atom of the sugar interacts with the purine- or pyrimidine base via a N-glycosidic bond and the C5 atom is linked to the phos-

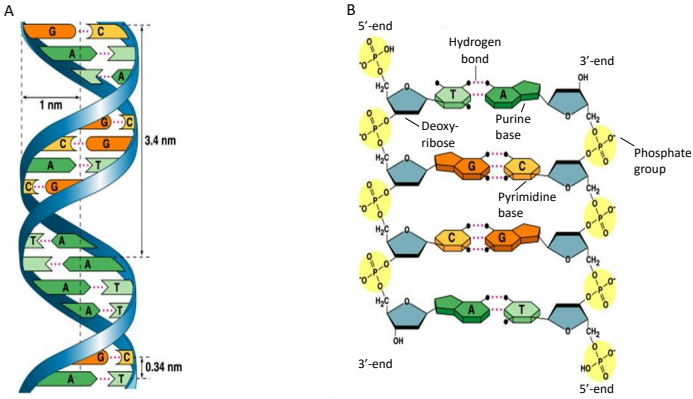


Figure 2.1: Structure of the DNA. (**A**) An illustration of the DNA double helix is shown. The blue lines represent the backbone of the DNA, while the bases adenine (A), guanine (G), cytosine (C) and thymine (T) are colored in dark green, light green, yellow and red, respectively. (**B**) A more detailed visualization of a short piece of double stranded DNA. The complementary bases interact via hydrogen bonds with each other. The backbone of the DNA consists of a deoxyribose (marked in blue) and a phosphate group (yellow circle). Figure obtained from Campbell and Reece (2009), Abb. 16.7 (a) and (b), p. 416, slightly modified.

phate group via an ester bond. Within the DNA, four different nucleotides are observed, which only differ in the base used. The bases of the DNA are the purine bases adenine (A) and guanine (G) and the pyrimidine bases cytosine (C) and thymine (T) (Nordheim and Knippers, 2018, p. 29,30).
The nucleotides are connected to each other such that they form a continuous non-branching strand. The DNA molecule consists of two antiparallel nucleotide strands which are wrapped around each other, forming the typical double helix structure. Thereby the sugar and the phosphate group build the backbone of the DNA, whereas the bases are directed towards each other such that they can interact and thereby stabilize the DNA structure (see Figure 2.1A). The base pairing follows specific rules: via hydrogen bonds adenine interacts with thymine and guanine with cytosine (see Figure 2.1B). A consequence of the base complementarity is that by knowing one strand, the other can be determined (Nordheim and Knippers, 2018, p. 30-32).
A second kind of nucleic acid is the ribonucleic acid (RNA), which consists of nucleotides formed by the same three components as the DNA. However, there are two major difference between RNA and DNA: The sugar of the RNA is ribose, and instead of the base thymine, the RNA contains uracil. In general, an RNA molecule is composed of only one strand. Nevertheless, RNA can exist in a double stranded form, where the base pairing occurs between

short complementary parts within one strand (Nordheim and Knippers, 2018, p.52). RNA molecules are typically classified into several classes. The messenger RNA (mRNA) is the best known one, and is translated into proteins. However, only a small fraction (around $1-2\%$) of the human genome encodes for genes that are translated to proteins (The ENCODE Project Consortium, 2012). The remaining genome is non-coding. A rather large fraction of the non-coding genome still results in functional RNAs. These non-coding RNAs (ncRNAs) can be involved in the biosynthesis of proteins like transfer RNAs (tRNA), be part of a subunit of a ribosome like ribosomal RNA (rRNA) or have regulatory effects like ncRNAs or small RNAs (sRNAs) (Nordheim and Knippers, 2018, p.52-54). Especially in recent years, it has been shown that ncRNAs are widely expressed and involved in gene regulation by controlling the chromatin architecture and enhancer activity (Mattick *et al.*, 2023). Additionally they play an important role in diseases, such as cancer (Slack and Chinnaiyan, 2019) and cardiovascular diseases (Gurha, 2019).

2.1.2 Proteins

Proteins are macromolecules comprised of amino acids. A single amino acid contains a carboxyl group, an amino group, a hydrogen atom and a side chain on the central C-atom. The side chain is different for each amino acid in terms of size, shape and charge. Among the existing amino acids, mostly 20 different amino acids are found in proteins (Campbell and Reece, 2009, p. 107-109). To build proteins these 20 amino acids can be put together in various combinations resulting in amino acid chains of different lengths (Nordheim and Knippers, 2018, p. 57). The amino acid chain(s) of a protein are folded in a characteristic shape resulting in a three-dimensional structure, which is essential for their functionality. One distinguishes between four different structural properties: The primary structure is the sequence of the amino acids (Nordheim and Knippers, 2018, p. 57), whereas the secondary structure is characterized by hydrogen bonds between CO- and NH- groups of nearby amino acids within the same chain. This results in three-dimensional sections like α-helices, β- sheets, β-loops or Ω-loops (Nordheim and Knippers, 2018, p. 60,61) (Berg *et al.*, 2013, p. 39-44). The arrangement of these three dimensional sections within the protein is called tertiary structure and is formed by disulphide bridges, hydrophobic interactions, *Van der Waals* forces and ionic bonds between far apart amino acids (within the linear sequence) (Nordheim and Knippers, 2018, p. 62,64), (Berg *et al.*, 2013, p. 45 - 48). A protein can consist of more than one amino acid chain, thereby the side chains of the amino acids of different chains can interact with each other. This spatial arrangement of the subunits of a protein is called quaternary structure (Nordheim and Knippers, 2018, 66) (Berg *et al.*, 2013, p. 47,48). Figure 2.2 summarizes the different structures. A well

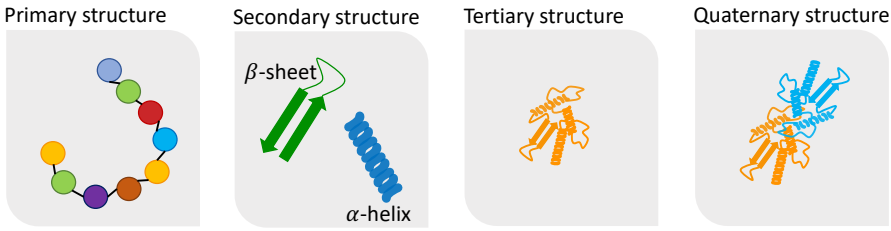


Figure 2.2: Protein folding. A visualization of the different states of protein folding. (from left to right) The primary structure is formed by the amino acid sequence, which can fold to small three dimensional sections like α-helices or β-sheets, called secondary structure. The tertiary structure is given by the arrangement of the secondary structure and the quaternary structures is the interaction of several folded amino acid chains.

known example is for instance the cro-protein of bacteriophages which consists of two identical subunits (Anderson *et al.*, 1981) or the the DNA polymerase α in human, which is built out of for different subunits (Lehman and Kaguni, 1989).

2.1.3 Definition of genes and genome

The word *gene* was first mentioned in 1909 by the danish botanist Johannsen (Johannsen, 1909). While the first definition of a gene was rather abstract, it became more and more detailed with further understanding. Around 1940 a popular definition was that one gene is a section of the DNA that produces one enzyme or polypeptide (Beadle and Tatum, 1941). However, in the following decades where the field of molecular genetics developed with rapid pace it became clear that the concept *gene* is much more complicated (reviewed for instanced in Portin and Wilkins (2017) and Gerstein *et al.* (2007)).
A current definition of genes is given by Petter Portin and Adam Wilkins introduced in their paper *The evolving definition of the term gene* (Portin and Wilkins, 2017). In their work a gene is defined as a not necessarily contiguous DNA region which results in one or more (non-)coding RNA molecules. The gene products interact which each other, forming so called gene regulatory networks. Depending on whether a gene's product participates as an element in a gene regulatory network or as an output, it can have indirect or more direct effects on the phenotype.
The exact number of human genes is still under discussion (Salzberg, 2018). The GENCODE annotation version 35 lists 19,954 protein-coding genes, 17,957 ncRNAs and 7,569 sRNAs (Frankish *et al.*, 2020).
The genome of an organism is described as the entire genetic information, in other words the DNA-sequence of all chromosomes. The human genome for

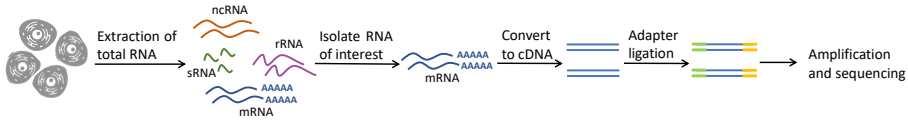


Figure 2.3: Overview RNA-seq. The different steps of an RNA-seq experiment for mRNA is visualized. Figure inspired by Kukurba and Montgomery (2015), Figure 1 and Nordheim and Knippers (2018) Abb. 26.9, p. 541.

instance consists of 24 chromosomes, counting in total around 3 billion base pairs (Berg *et al.*, 2013, p. 163).

2.1.4 Gene expression measurement

The genes encoded in the DNA are transcribed into RNA. Transcription is the synthesis of RNA molecules, where the DNA is the blueprint and the nucleotides of the DNA are rewritten (or transcribed) into a nucleotide sequence of RNA.
To measure gene expression in bulk data, the standard technique is RNA-seq (Wang *et al.*, 2009). In a first step, RNA is extracted from the cells and the RNA of interest is isolated, such as the total amount of RNA, only the mRNA (mRNA enrichment), or everything beside rRNA (Ribo-depletion). The resulting RNA is fragmented and converted to cDNA, which is the complementary DNA of an RNA fragment. Next, adapters are ligated to the fragments which are then amplified and sequenced (Kukurba and Montgomery, 2015; Stark *et al.*, 2019). The steps are summarized in Figure 2.3. Computational steps are necessary to quantify the expression of the genes, and are outlined in Section 2.2.2.

2.1.5 Chromatin structure and accessibility

In eukaryotic cells, the DNA is packed into chromatin. Beside the DNA-sequence itself, the chromatin mostly consists of different kinds of histones (Nordheim and Knippers, 2018, p. 149). A histone is a basic/alkaline protein in which around a quarter of all amino acids consist of lysine or arginine (Berg *et al.*, 2013, p. 951). The protein sequence of histones is folded to three α-helices, which are connect to each other with short loops (Nordheim and Knippers, 2018, p. 147). Among the existing variants of histones, most common are H2A, H2B, H3 and H4, which can form the protein complex found in chromatin. The resulting histone octamer consists of eight histones, where H2A, H2B, H3 and H4 each occur twice. A 147 bp long DNA sequence is wrapped around a histone octamer forming the so-called nucleosome, which are separated from each

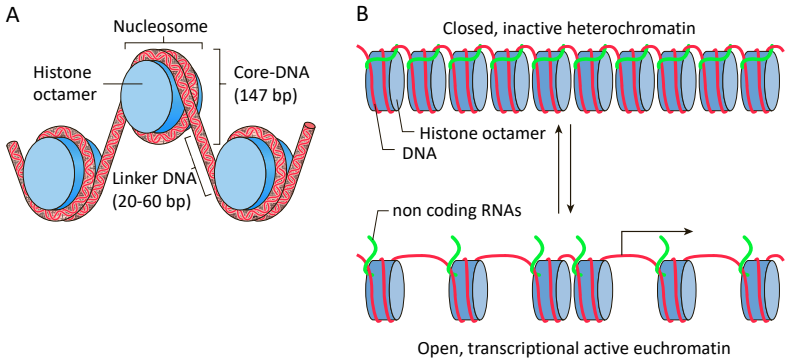


Figure 2.4: Chromatin structure. (**A**) DNA wrapped around histone octamers forming nucleosomes, which are connected which each other through linker DNA. Figure obtained from Nordheim and Knippers (2018), Abb. 7.10, p. 149 modified. (**B**) Visualization of densely packed, inactive heterochromatin (top) and loosely packed, active euchromatin (bottom). Epigenetic modifications can transform one state into the other. Figure obtained from Nordheim and Knippers (2018), Abb. 20.2, p. 444 modified.

other by a linker DNA (see Figure 2.4A) (Nordheim and Knippers, 2018, p. 149) (Richmond and Davey, 2003). In general one can distinguish heterochromatin, which is densely packed and often associated with inactive regions. A popular example for heterochromatin is the inactivation of one X-chromosome in female mammals (Lyon, 1961). In contrast, there is euchromatin, which is more loosely packed to allow transcriptional activity (Nordheim and Knippers, 2018, p. 147) (see also Figure 2.4B).
The positioning of the nucleosomes in the genome is dynamic and is involved in DNA-dependent processes like transcription, DNA-repair or replication. Further, it can have a regulatory function by altering the accessible regions to which DNA-binding proteins can bind to (Tsompana and Buck, 2014). Additionally, it has been shown that mutations in chromatin remodelers impact the structure of the chromatin and lead to consequences for human health (Gaspar-Maia *et al.*, 2009; Hargreaves and Crabtree, 2011; Schwartzentruber *et al.*, 2012).
Several techniques exist to study the genome-wide chromatin accessibility (reviewed for instance in Tsompana and Buck (2014) or Klein and Hainer (2019)). There is MNase-seq that indirectly measures open chromatin by degrading open chromatin regions with MNase digestion. The remaining regions occupied by nucleosomes are sequenced. In contrast, there are methods that directly detect open chromatin regions. For instance, DNase1-seq utilizes the enzyme DNase I that has the ability to cut nucleosome free regions, so called hypersensitive sites. After DNA purification, short DNA-fragments, which arise

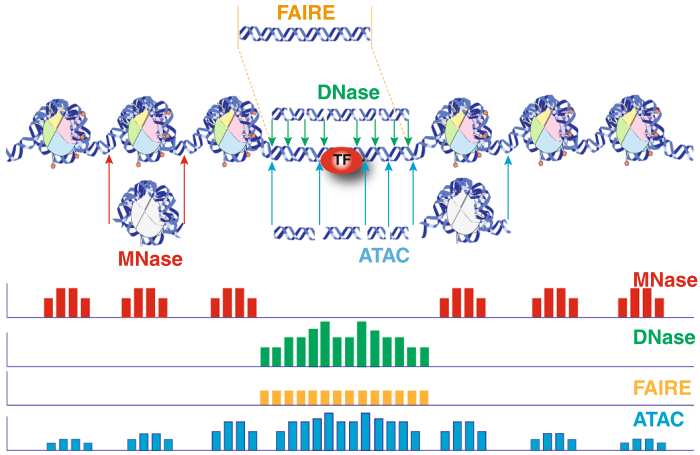


Figure 2.5: Schematic representation of techniques that measure open chromatin. (top) Visualization of a short region of open chromatin flanked by nucleosomes. The arrows point to the DNA fragments generated by the different techniques indicated by different colors. (bottom) Each row shows the resulting signal of a technique. The higher the bar the more signal was obtained. In the signal of DNase1-seq and ATAC-seq footprints can be observed, which indicate the location of DNA-binding proteins, like Transcription Factors (TF). Figure is taken from Tsompana and Buck (2014), Figure 1.

through high cleavage activity of DNase I, are extracted and sequenced. An alternative technique is formaldehyde-assisted isolation of regulatory elements (FAIRE)-seq, where proteins are cross-linked to the DNA. Through sonication the DNA is sheared, resulting in fragments bound by proteins and protein-free fragments, which are filtered out and sequenced. A rather new technique to identify open chromatin is the assay for transposase accessibility and deep sequencing (ATAC-seq), which relies on a hyperactive Tn5 transposase. The enzyme inserts sequence adapters into open chromatin regions. The tagged regions are then isolated and sequenced. Figure 2.5 schematically outlines the resulting signal of the described techniques. In the signal track of DNase1-seq and ATAC-seq one can observe short regions, where the signal is slightly dropping. These drops in read coverage are called footprints. They originate from proteins that bind to the DNA and thus prevent the cleavage enzymes to cut at their position. Based on the characteristics of a footprint it is possible to predict which DNA-binding protein was observed. The sequenced DNA-fragments are mapped to the genome and further analyzed with bioinformatic techniques, outlined in Section 2.2.2.

2.1.6 Epigenetic modifications

The word epigenetics can be translated as *above genetics*, and is a field in molecular biology where mechanisms are studied that lead to heritable changes in the structure and activity of the chromatin without affecting the DNA sequence itself (Dupont *et al.*, 2009). A well established epigenetic modification is the DNA methylation, where the cytosine in a CpG context is methylated (for instance reviewed in Greenberg and Bourc'his (2019)). In the following, we explain histone modifications, experimental techniques for their genome-wide identification and their regulatory impact.

Histone Modifications

The N-terminal regions of the protein sequence of histones can be posttranslationally modified, which has not only an impact on how the histones interact with the DNA (Kouzarides, 2007), but also allows the recruitment of protein complexes, mostly complexes that modify the chromatin (Nordheim and Knippers, 2018, p. 151). Among the most studied modifications of the side chains of histones are acetylation and methylation. Special enzymes, the acetyltransferases (HATs) can transfer an acetyl-group to the lysine (K) side chain of the histones H2A, H2B, H3 and H4. This is a reversible process, since a histone deacetylase (HADC) can remove the acetyl-group again. Similarly, the side chains of the amino acids lysine and arginine (R) of H3 and H4 can be methylated by a histone methyltransferase or demethylated by a histone demethylase (Nordheim and Knippers, 2018, p. 151, 152).
To abbreviate the different histone modifications, we follow the nomenclature in Nordheim and Knippers (2018): First the histone is denoted, followed by the modified amino acid and the kind of modification. For instance, H3K4me3 is the abbreviation for a modification at histone H3, where the amino acid lysine at position 4 of the side chain is tri-methylated.
Specific histone modifications are often found at the same functional sites in the genome and certain modifications co-occur with other proteins or transcriptional events. For example, at the TSS of highly expressed genes a high level of RNA polymerase II was detected, strongly correlated with H3K4me3. Also H3K4me2 and H3K4me1 signals are often detected around transcription start sites (TSSs) of active genes (Barski *et al.*, 2007). Further, H3K36me3 is a modification associated with elongation and mostly found in transcribed regions (Bannister *et al.*, 2005). In contrast, H3K27me3 or H3K9me3 are associated with heterochromatin and therefore with inactivated and repressed genes (Jiang and Mortazavi, 2018). In Table 2.1 the most commonly studied histone modifications are summarized and in Figure 2.6 the signal tracks of some of these modifications in relation to gene activity at the TSS are shown.
Histone modifications are maintained through DNA replication, however, the

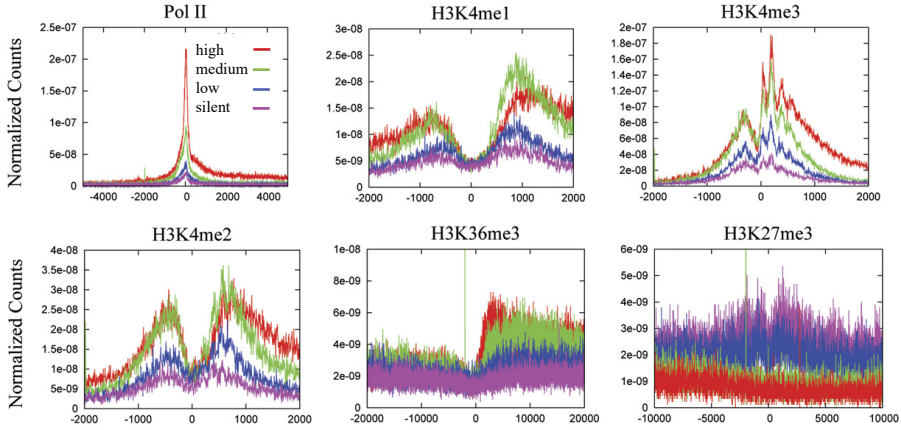


Figure 2.6: Histone modifications of the TSS depending on the genes' expression. For six histone marks the signal is visualized around the TSS categorized based on the gene expression strength (color coding). Figure is taken from Barski *et al.* (2007), Figure 2 A,B,C,D,E,G, adapted.

exact mechanism is currently not entirely understood (Francis and Sihou, 2021; Budhavarapu *et al.*, 2013)

Histone modification	primary localization and functionality
H3K4me3	active promoter
H3K4me1	active and poised enhancers
H3K27ac	active enhancers
H3K36me3	actively transcribed genes
H3K9me3	heterochromatin
H3K27me3	heterochromatin

Table 2.1: Commonly studied histone modifications. The table outlines 6 different histone modifications, their primary location and their associated functionality.

Genome-wide Identification of Histone Modification

A commonly usedtechnique to detect histone modifications, the location of nucleosomes and also DNA binding proteins in a genome-wide fashion and at base pair resolution is chromatin immunopercipitation followed by sequencing (ChIP-seq) (Park, 2009). For a histone ChIP-seq analysis, the cells can

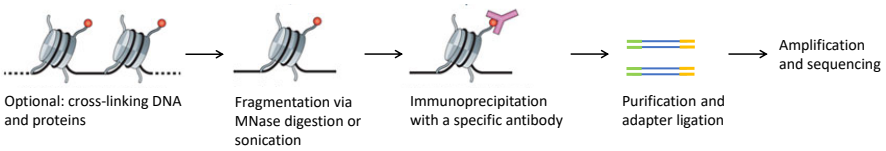


Figure 2.7: Overview of a histone ChIP-seq experiment. Illustration of the steps of a histone ChIP-seq experiment. The studied histone modification is marked with a red dot on one of the side chains of a histone. Parts of the figure are taken from Furey (2012), Figure 1(b).

be treated with formaldehyde to cross-link the nucleosomes to the DNA followed by fragmentation with sonication (Orlando, 2000). It is also possible to omit cross-linking and directly digest the DNA with an micrococcal nuclease (MNase) resulting in chromatin fragments (O'Neill, 2003). Both techniques are commonly used und have their own advantages and disadvantages (Park, 2009). Next, the histone of interest is marked with a specific antibody by immunoprecipitation, and is then tagged with antibodies to gather the attached fragments. The fragments are purified and if necessary the cross-linking is reversed, such that they can be sequenced. To distinguish enriched fragments from those which result from side effects of the experiment, the analysis is repeated without the antibody, known as input (Park, 2009; Furey, 2012). These steps are also summarized in Figure 2.7. Since histone modifications are dynamically placed and removed in the genome and are cell type-specific, a ChIP-seq experiment results in a snapshot of a specific timepoint averaged across the analyzed cells. The resulting reads from the input and the ChIP-seq experiment with antibody are aligned to the reference genome using tools like bowtie2 (Langmead and Salzberg, 2012). To detect the locations where the analyzed histone mark occurs within the genome, a peak caller like MACS2 (Zhang *et al.*, 2008) or SICER2 (Zhang and Lupski, 2015) is used to identify enriched regions (see Section 2.2.2).

The ChIP-seq technique is established as a key method in the epigenetic field and contributed to important findings in transcriptional regulation (reviewed in Nakato and Shirahige (2016)). However, there are a few drawbacks of the ChIP-seq technique which one needs to keep in mind. The quality of the ChIP-seq results strongly depends on the quality of the used antibody (Park, 2009). The more sensitive the antibody is the more specific is the immunoprecipitation step. The quality between the antibodies are highly variable, and need to be validated in the lab. It can occur that no reliable antibody for the studied modification or DNA-binding protein is known, so the experiment can not be performed.

2.1.7 Gene regulation

To understand how genes are regulated, one aims to read the regulatory information encrypted within the DNA sequence by decoding the so-called cis-regulatory code. Even though this is still an unsolved problem, and one is not able to read the cis-regulatory code universally throughout the genome, a lot is known about gene regulation and the components involved. In the following, TFs are introduced and their interplay with regulatory elements is outlined.

Transcription factors

Transcription Factors (TFs) are proteins that directly interact with the DNA by recognizing short patterns in the sequence to which they can bind to. TFs control the regulation of gene transcription by forming complexes with other proteins as for instance with the RNA polymerase (Reiter *et al.*, 2017). They are also known to be involved in directly affecting the chromatin state, for instance by acting as pioneer factors (Zaret, 2018). Further, TFs play an important role in countless functional processes (Vaquerizas *et al.*, 2009; Lambert *et al.*, 2018). For example, TFs are involved in cell cyle maintenance (Simon *et al.*, 2001; Dynlacht, 1997), in the differentiation to specific cells types (Bain, 1994; Accili and Arden, 2004; Flitsch *et al.*, 2020), in germ-cell development (Kojima *et al.*, 2021) and in inducing apoptosis (Gao *et al.*, 2023). Further, TFs are often deregulated in cancer (Islam *et al.*, 2021; Bhagwat and Vakoc, 2015) and it has been shown that changes in TF binding sites are linked to differential gene expression between individuals (Kasowski *et al.*, 2010) and can also cause diseases (Deplancke *et al.*, 2016). TFs can enhance the transcription of a gene but also repress it (Ma, 2005), they can do so for multiple different genes (Verheul *et al.*, 2020) and their function is cell type-specific (Lee *et al.*, 2011; Vaquerizas *et al.*, 2009) Another interesting observation Vaquerizas and colleagues point out, is that the expression of TFs is significantly lower compared to proteins not encoding for a TF across several cell types (Vaquerizas *et al.*, 2009).
TFs are equipped with a DNA binding domain (DBD), which is a tertiary protein structure that enables them to directly interact with the DNA-sequence (Luscombe *et al.*, 2000). Based on the DBD, TFs can be grouped in a hierarchical system to various kinds of classes and families, as it is done for instance in the TFClass system (Wingender *et al.*, 2017). The different DBDs allow conclusions about the general functionality of the TFs, their structure and their binding characteristics. For instance, the TFs with a *Homeo domain factors* DBD are often involved in development and cell differentiation. Some of the TFs with this DBD are observed to be widely expressed, whereas others are cell type-specific. Most TFs that belong to the *Homeo domain factors* DBD are rather small proteins consisting of four helices, which either bind as monomer or dimer to the DNA (Luscombe *et al.*, 2000) (see Figure 2.8A). Interestingly,

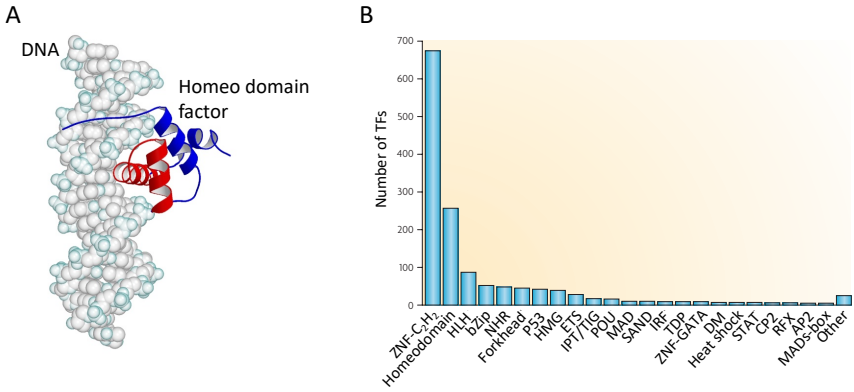


Figure 2.8: DNA-binding domains of TFs. (**A**) The 3 dimensional structure of a TF with the DNA-binding domain *Homeo domain factors* interacting with the DNA is shown. The DBD is depicted in red, and the remaining structures of the TF is colored in blue. (**B**) A histogram shows how many TFs belong to each group of DBDs based on InterPro database (Hunter *et al.*, 2009). By far the most known TFs are *C2H2 zinc finger factors* (ZNF-C_2H_2), followed by *Homeo domain factors* (Homeodomain) and *Basic helix-loop-helix factors* (HLH). Part A is taken from Luscombe *et al.* (2000), Figure 1(b), slightly modified and Part B from Vaquerizas *et al.* (2009), Figure 2, increased font.

most of the TFs in human and mouse belong to only three DBD classes, the *C2H2 zinc finger factors*, the *Homeo domain factors* and the *Basic helix-loop-helix factors* (see Figure 2.8B) (Vaquerizas *et al.*, 2009; Gray *et al.*, 2004).
To understand how TFs regulate gene transcription, it is necessary to localize the TFs binding sites within the genome. Many experimental techniques are available to determine the binding sites of a TF either *in vivo* or *in vitro* (summarized in *e.g.* Lambert *et al.* (2018)). For example a ChIP-seq targeting TFs can be used similarly as explained in Section 2.1.6 for histones. The computational details for further analyzing the resulting reads are outlined in Section 2.2.2. TFs might also act as co-factors and interact with other TFs, where they do not directly bind the DNA. However, the antibody used within the ChIP-seq experiment will still identify these cases, resulting in DNA fragments bound by the co-factors. Thus, binding sites of co-factors of the studied TFs might be included in the resulting peak regions.

Regulatory elements

Regulatory elements (REM) are non-coding DNA regions to which TFs can bind and by that affect the transcription of a gene. REMs can for instance be promoters, which are non-coding DNA regions directly upstream of a gene that

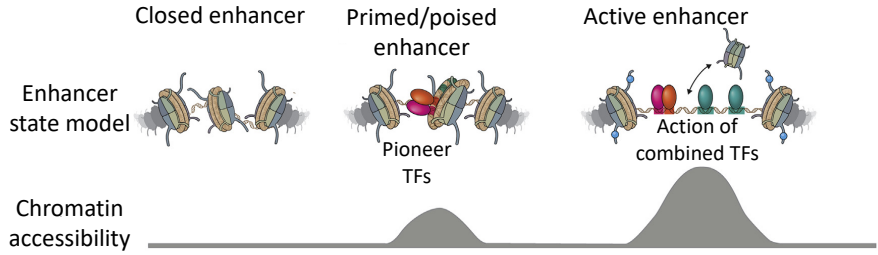


Figure 2.9: Activation of enhancers. An illustration that shows how an enhancer is activated (top) with respect to the consequence for the chromatin status (bottom). (left) A closed chromatin state is shown. (middle) Pioneer TFs bind to the closed chromatin to make the region accessible. (right) More TFs are recruited, histone modifications associated with active enhancers are set (marked by blue circles) and the central histone is removed. Figure is taken from Zeitlinger (2020), Figure 3, slightly increased font.

initiate the transcription and include its transcription start site (Weingarten-Gabbay *et al.*, 2019). Another type of REMs are enhancers that are typically hundreds of base pairs long and can be located distally from the regulated gene (Li and Wunderlich, 2017). Some enhancers are found to be more than a million bp away from their target gene, as for instance an enhancer for the *SHH* gene (Lettice, 2003). The chromatin content and also the cell type-specific TFs affect the functionality of an enhancer (Maricque *et al.*, 2018; Heinz *et al.*, 2015). In the process of activating an enhancer many TFs and other proteins can be involved. Pioneer TFs bind to closed DNA regions and make the region accessible (Zaret, 2018) by recruiting co-factors, which further promote chromatin accessibility (Zeitlinger, 2020). Additionally, histone-modificating enzymes can be recruited, among them a histone methyltransferases, histone acetyltransferase or histone deacetylases, which modify the tails of enhancer-associated nucleosomes (Heinz *et al.*, 2015). Further, chromatin remodelers can move nucleosomes along the DNA sequence (Spitz and Furlong, 2012). Active enhancers have displaced the central nucleosome to further increase the chromatin accessibility (He *et al.*, 2010) (see Figure 2.9).
In a common model, active enhancers are bound by TFs and ensure the interaction to the gene's promoter through a 3D-loop to support transcription (see Figure 2.10) (Gasperini *et al.*, 2020). Beside the 3D-loop, enhancers can interact with promoters in different ways, described with the tracking or linking model (reviewed in Furlong and Levine (2018)). Additionally, chromatin-capture and imaging data suggest much more complex interactions, where multiple enhancers and promoters are organized in topologies (Rao *et al.*, 2014).
Based on the epigenetic signature of enhancers it is possible to predict the enhancer status within the studied cell type (Ernst and Kellis, 2010). Inac-

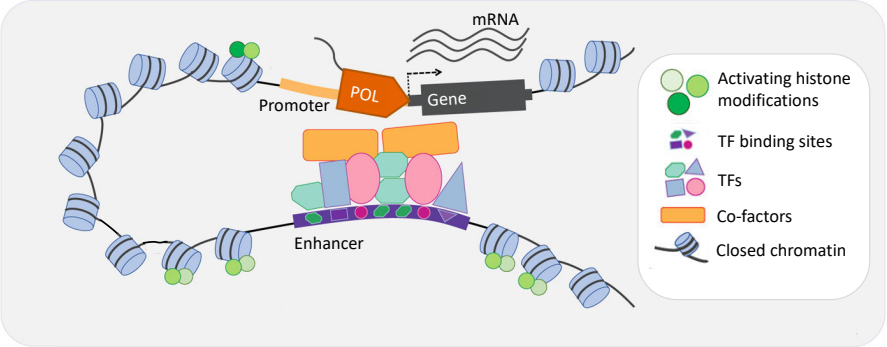


Figure 2.10: Enhancer-promoter interaction. An active enhancer that interact with a promoter via a 3D-loop is visualized. The enhancer is bound by multiple TFs and co-factors, and the genomic region around the enhancer is marked with activating histone modification. Figure is taken from Gasperini *et al.* (2020), Figure 1A, slightly modified.

tive enhancers are characterized by closed chromatin and usually marked with H3K27me3. Active enhancers are open chromatin regions that show an enrichment of H3K4me1 or H3K4me2, and a less strong signal of H3K4me3 than promoters (Heintzman *et al.*, 2007). Further, active enhancers are often marked by H3K27ac (Creyghton *et al.*, 2010). Another characteristic is that they are bi-directionally transcribed, resulting in enhancer RNA (eRNA) (Blackwood and Kadonaga, 1998; Mikhaylichenko *et al.*, 2018; Santa *et al.*, 2010). Additionally, there are primed enhancers, which are already bound by TFs that established open chromatin, and are marked with H3K4me1-2, but no eRNA is produced (Spicuglia and Vanhille, 2012). Poised enhancers are marked with H3K27me3, while H3K4ac is absent. They are associated to inactive genes in early development and are frequently found in human embryonic stem cells (Rada-Iglesias *et al.*, 2010). In Figure 2.11 the epigenetic characteristics of different chromatin states are compared.
Open chromatin data, histone and TF ChIP-seq, as well as Hi-C data can be used to predict enhancers genome-wide, as it was shown by various kinds of studies (Visel *et al.*, 2009; Andersson *et al.*, 2014; Gao and Qian, 2019b; Schmidt *et al.*, 2021). Additionally, measurements like massively parallel report assays (MPRA) allow to test the functional activity of thousands of candidate REMs in one experiment and thus can be used to identify enhancers (de Almeida *et al.*, 2022; Cai *et al.*, 2019).

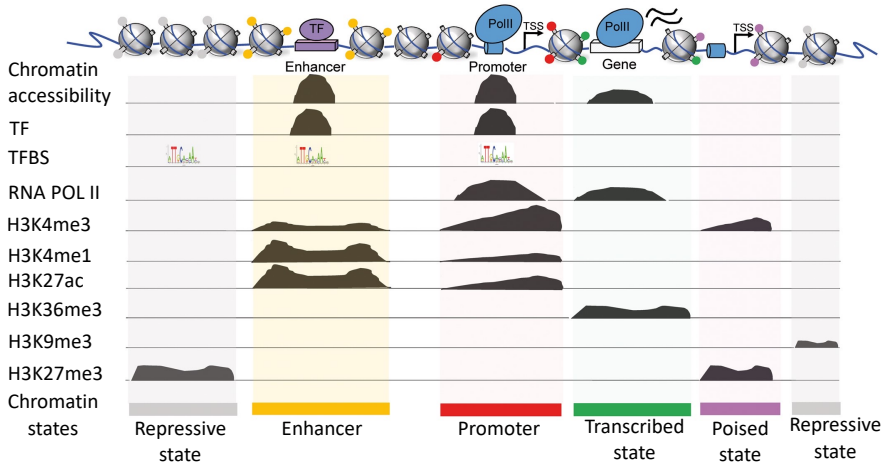


Figure 2.11: Comparison of histone modifications of different chromatin states. An illustration that visualizes TF, RNA polymerase II and histone ChIP-seq signals and chromatin accessibility data (rows) for a genomic region with differential functional sites (columns). The different chromatin states can be characterized by a typical epigenetic signature. Figure is taken from Jiang and Mortazavi (2018), Figure 1, increased font.

2.1.8 Functional genomics data in public resources

Large consortia, such as ENCODE (The ENCODE Project Consortium, 2012), Blueprint (Stunnenberg *et al.*, 2016) or Roadmap (Kundaje *et al.*, 2015) provide uniformly processed functional genomics data *e.g.* open chromatin, ChIP-seq or RNA-seq data. The ENCODE consortium analyzed hundreds of ChIP-seq experiments for various kinds of TFs and histone modifications, RNA-seq, as well as chromatin accessibility in cell types or tissues mostly from human and mouse. Currently 12,829 samples for human are available (https://www.encodeproject.org; April, 2023). They ensure standardized data processing with uniform pipelines, which are freely available. These efforts result in high quality data which are publicly available for the community and are incorporated in more than 2,000 studies from researchers not part of the ENCODE consortium (The ENCODE Project Consortium., 2020).
In comparison to ENCODE which is still ongoing, Blueprint and Roadmap are already completed projects. The Blueprint consortium is specialized to human haemopoetic epigenomes including RNA-seq, histone ChIP-seq and DNase1-seq for cell types or tissues, such as monocytes, macrophages or venous blood, as well as disease-associated samples. In contrast, the Roadmap database holds 111 human epigenomics data sets of a broad range of cell types and tissues, such as heart, lung or muscle but also embryonic stem cells and induced pluripotent

stem cells.
In this thesis publicly available ChIP-seq, open chromatin and expression data from these resources are used.

2.1.9 Genetic variations and their measurement

A topic of interest is to understand the effect of genetic variation, especially in the non-coding DNA regions. In the following, we first outline the different mutations that can occur in the genome. Thereafter, we describe how genetic variations are detected and associated to diseases and traits. Finally, we summarize the impact of mutations on the coding and non-coding genome.

Basic definition of mutations

Mutations are genetic variations of the genome, that are caused by errors during DNA-replication or by recombination, but also by environmental effects like sonication or chemicals (Campbell and Reece, 2009, p. 461-462). Mutations drive evolution, since they create variability between individuals which makes evolutionary changes possible (Nordheim and Knippers, 2018, p. 250). Mutations are mostly studied in the context of diseases, but they can also have positive effects.
Several types of mutations are observed, some mutations can affect the number, the structure or the form of a chromosome. Polyploidy, for example, describes the gain of a whole copy of the genome. Translocation is the process where a chromosomal region is relocated either within the chromosome or to another one. Inversions, on the other hand, is the change of orientation of pieces of DNA (Loewe and Hill, 2010). Another type of well-known mutations are insertions or deletions, so called InDels of one base pair or a rather short genomic region. Further, there a mutations where one base pair is exchanged by another one, denoted as Single Nucleotide Polymorphism (SNP). In Figure 2.12 an example for an insertion, a deletion and a SNP is outlined. Mutations can occur in coding, as well as in non-coding regions of the genome (Nordheim and Knippers, 2018, p. 250-253). InDels and SNPs result in two different allele variants of the same genomic locus. The common one is usually denoted as wildtype allele, whereas the less common one is called alternative allele. It is also possible to observe more than one alternative allele.

Experimental techniques to detect genetic variations

By genotyping one can identify genetic variations of individuals or populations based for instance on whole-genome sequencing (WGS), as it is done in a large scale in the 1000 Genomes project (The 1000 Genomes Project Consortium,

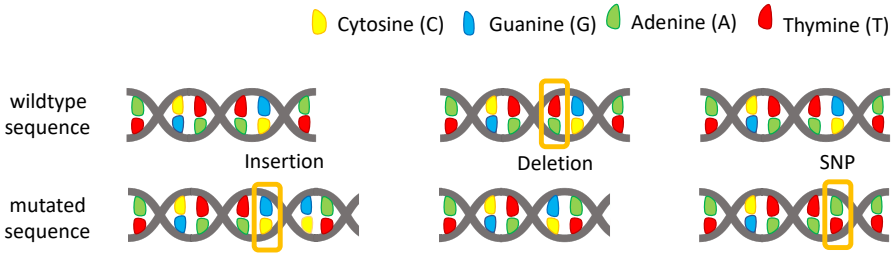


Figure 2.12: Genetic variations. An illustration visualizing an example for an insertion, deletion and a SNP. (top) A short wildtype sequence is depicted, where the bases are represented by shapes in yellow, green, blue and red representing cytosine, adenine, guanine and thymine, respectively. (bottom) Different genetic variants with respect to the wildtype sequence are outlined . (left) A guanine paired to cytosine is inserted between the fourth and fifth position. (middle) The adenine-thymine base pair at the fourth position is deleted. (right) At the fifth position an adenine paired to a thymine is exchanged by a guanine-cytosine base pair. The genetic variation is highlighted with an orange box.

2010). The aim was to identify common genetic variants, which have a minor allele frequency (MAF) $\geq 1\%$ in the studied human populations. For a genetic variation the MAF describes the frequency of the second most common alternative allele observed in a population. SNPs that are not common are denoted as rare SNPs (usually with a MAF $\leq 1\%$) (Gibson, 2012). Additionally, the authors of the 1000 genome project aimed to provide haplotype information of the studied populations. Haplotypes are alleles which are usually inherited together, since they occur in close vicinity to each other such that the separation by recombination is unlikely (Nordheim and Knippers, 2018, p. 480). he term linkage disequilibrium (LD) is used to express the degree of how frequently several alleles are inherited together within a population (Nordheim and Knippers, 2018, p. 510).

Genetic variations are collected independently of their MAF and functionality in the dbSNP database (Sherry, 2001). To keep the database up-to-date, research groups can easily submit newly discovered genetic variations. The current release of the dbSNP database (build 155) holds $1,053,623,523$ genetic variations.

To associate genetic variations, such as SNPs and InDels to diseases or traits, a widely used technique is to perform a genome wide association study (GWAS). In a GWAS, differences in the allele frequency between individuals from the same population are analyzed with respect to the studied phenotype. The analysis can be separated into multiple steps, which are briefly outlined in the following based on the recent review from Uffelmann *et al.* (2021). In a first step, the data of the individuals with the phenotype of interest and a control

group needs to be collected either based on a population or a family. Often the data is combined with publicly available resources, such as UK Biobank (Fry *et al.*, 2017) or Biobank Japan (Nagai *et al.*, 2017). Next, the considered individuals are genotyped. Depending on the aim of the study either microarrays are used to detect common genetic variations or WGS which allows to include rare SNPs as well. Thereafter, a quality control is applied which is usually done with tools explicitly developed for this purpose, such as PLINK (Purcell *et al.*, 2007). The detected genetic variants undergo phasing, which estimates whether the observed allele is inherited from a maternal or paternal allele. Doing so one can identify haplotypes, which are inherited together. Based on publicly available haplotype reference panels *e.g.* from the 1000 Genomes project, missing SNPs not genotyped can be imputed. After the data processing, an association test is used to detect genetic variations linked to the studied phenotype. Depending on the phenotype different statistical models are used, for instance for a continuous phenotype like blood pressure usually a linear regression is applied. Since millions of associations between the genetic variations with respect to the studied phenotype are tested, it is necessary to perform multiple testing correction and to choose a stringent cutoff to avoid false positives. To improve the performance and the reliability of GWAS multiple smaller cohorts can be combined in a meta-analysis, and the results are replicated either internally or with an external cohort. A GWAS results in a summary statistic, from which the significantly associated SNPs can be gathered. However, how exactly these SNPs affect the phenotype is not clear and several experimental or computational approaches for follow-up analyses can be applied.
Thousands of GWAS have been performed for a wide range of traits and diseases, such as coronary artery disease (Nelson *et al.*, 2017), autoimmune hepatitis (Alberts *et al.*, 2017) or pulmonary fibrosis (Allen *et al.*, 2020), but also for body height at birth (van der Valk *et al.*, 2014) or eye color (Rawofi *et al.*, 2017). The NHGRI-EBI GWAS Catalog (Buniello *et al.*, 2018) is a public database holding the summary statistics of 5687 GWAS from $3,567$ publications resulting in more than $71,000$ variants associated to traits or diseases.
A helpful resource to pinpoint to causal non-coding SNPs is the Genotype-Tissue Expression (GTEx) database (GTEx Consortium, 2017), which holds expression Quantitative Trait Loci (eQTLs) for 49 tissues and cell types. An eQTL is a locus in the genome, mostly a genetic variation, such as a SNP, that modulates gene expression (Vandiedonck, 2018; Nica and Dermitzakis, 2013). Thus, eQTLs are linked to target genes. In an eQTL analysis an association test is used to detect the changes between the different forms of the genetic variation, in case of a SNP the different alleles, in relation to gene expression in a large number of samples (see Figure 2.13). eQTLs can be combined with GWAS to identify SNPs which are not only linked to a gene and affect its expression, but also to the studied phenotype (Uffelmann *et al.*, 2021).

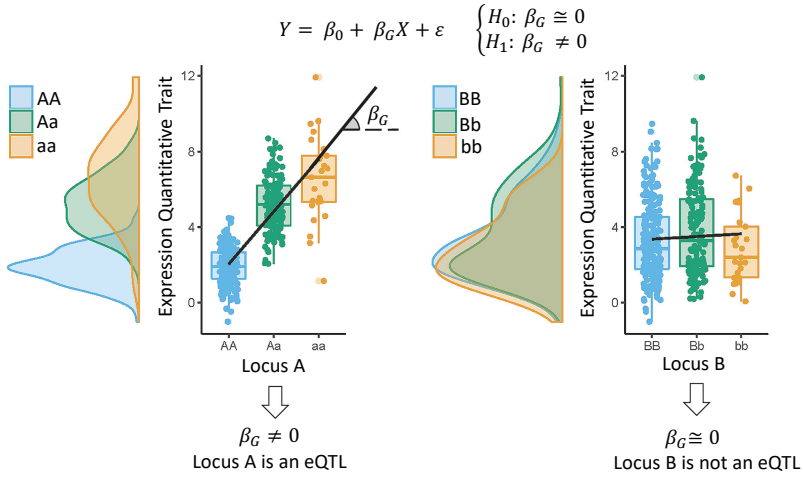


Figure 2.13: eQTLs detection. The loci A and B are analyzed to determine whether they are eQTLs or not. In each panel the expression of the studied genetic variation for hundreds of samples is shown with respect to the analyzed genotypes either as density or boxplot. To test for an association usually a logistic regression is used, where Y holds the expression values of the target genes for all analyzed samples (phenotype) and X represents the genotype per sample. If the slope β_G is significantly different from 0, an eQTL is detects (left panel). The figure is taken from Vandiedonck (2018), Figure 1, slightly modified.

Consequences of coding and non-coding genetic variations

When a SNP or InDel occurs in a protein-coding region of the genome, it can have a direct effect on the resulting protein. A prominent example is a mutation in the β-globin gene, which causes sickle-cell anemia, a disease which negatively affects the oxygen transport (Campbell and Reece, 2009, p.372, 460). The exchange from T to A leads to a different amino acid in the resulting protein, which has an effect on the protein structure.
The ENCODE consortium analyzed trait- or disease-associated SNPs from the GWAS Catalog and noticed that 88% of these SNPs where located in introns or intergenic regions, thus they are non-coding SNPs. Out of the non-coding SNPs associated to a disease or trait, 12% overlapped with TF ChIP-seq binding sites and 34% overlapped with DNase hypersensitive sites (The ENCODE Project Consortium, 2012). As a conclusion, they may have a regulatory function. Several studies detected the effect of non-coding SNPs on gene regulation (review in Deplancke *et al.* (2016) or Zhang and Lupski (2015)). For instance, a SNP within the promoter of gene *HBG1* was shown to create a binding site of the TFs TAL1 and GATA1, and is associated with increased fetal globin (Wienert

et al., 2015). Another example is that a binding site of TF SIX3 in a regulatory element is interrupted by a SNP, leading to expression changes of *SHH* which is around 460kb away from the mutation (Jeong *et al.*, 2008). Besides of affecting binding sites of TFs within regulatory elements, it was shown that non-coding SNPs in cancer can also disrupt chromatin domain structures (Hnisz *et al.*, 2016), or occur in 5' or 3' untranslated regions and thereby affect mRNA translation and stability (Schuster and Hsieh, 2019).
Overall, it is still challenging to identify the impact of non-coding SNPs, also because the cis-regulatory code is not yet fully understood.

2.1.10 CRISPR-Cas as genome editing system

In recent years, the clustered regularly interspaced short palindromic repeats (CRISPR) and its associated protein Cas have become the most efficient and accurate technique for genome editing (Mengstie and Wondimu, 2021). The CRISPR-Cas system was discovered in bacteria and archaea and equips them with an adaptive immunity against viruses and plasmids by cleaving invasive DNA-sequences (Barrangou *et al.*, 2007). To become functional a CRISPR-Cas system relies on a CRISPR RNA (crRNA), which is usually denoted as guide RNA (gRNA) in an experimental setting. This gRNA specifies the target DNA sequence which is cleaved by the CRISPR-Cas system. The target sequence must be located next to a specific motif, which is responsible for initiating the binding between the target DNA and the gRNA. Different Cas proteins are known, among them Streptococcus pyogenes Cas9 which was the first one used in genome editing (Jinek *et al.*, 2012) and is still widely applied.
The CRISPR-Cas induced double strand breaks trigger the DNA damage response which repairs the DNA using various kinds of mechanisms (Wyman and Kanaar, 2006). Depending on the repair mechanisms used and the editing strategy, one can induce small insertions or deletions, as well as cut out large regions even of the size of megabases (Essletzbichler *et al.*, 2014). Doing so, one can for instance knock out genes or cut regulatory elements out of the genome (Pickar-Oliver and Gersbach, 2019). Further, it is possible to integrate a DNA fragment at the position of the cleavage site.
The different strategies, as well as further applications of the CRISPR-Cas system, are summarized in Pickar-Oliver and Gersbach (2019).

2.2 Mathematical and computational basics

In the following section, we outline basic statistical concepts used throughout this thesis, followed by a summary of bioinformatics tools to analyze sequencing data. Thereafter, it is explained how TF binding prediction is done, commonly

used techniques for TF motif enrichment are introduced and a cluster algorithm for TF motifs is described. Further, we provide a section that summarizes a method to identify regulatory elements and their linked target genes. Next, we describe what regulatory SNPs are, how they can be detected and how SNPs are associated to target genes. In the last section, we introduce the DisGeNET database.

2.2.1 Statistical basics

The following section outlines statistical measurements, concepts and distributions we used throughout this thesis, such as precision, recall and mean squared error but also random variables, the Laplace distribution and hypothesis testing.

Performance evaluation of classification tasks

In a classification the task it to predict to which category an observation belongs to. In order to evaluate the quality of a classifier, its performance on a test dataset is analyzed. The confusion matrix collects the number of observations belonging to the four possible outcomes: *True positives* (TP), *true negatives* (TN), *false positive* (FP) and *false negatives* (FN) (James *et al.*, 2013, 145-149). The dataset can be balanced or imbalanced, where the later is mostly used throughout the thesis, thus we focus on them and their evaluation. In an imbalanced dataset, the set of negatively and positively labeled observations is not equally in size. Commonly used measurements to evaluate imbalanced datasets are *precision* (P) and *recall* (R) (Saito and Rehmsmeier, 2015).
Given the observations that are predicted to the positive category, precision is the fraction of the correct predicted observations within this category:

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

Recall is the proportion of the fraction of positive labeled observations that are correctly predicted:

$$R = \frac{TP}{TP + FN} \tag{2.2}$$

Often the classification is dependent on an threshold t on which basis it is decided to which category an observation is predicted to. Varying this threshold, also affected the number of the four possible outcomes and thus the measurements precision and recall. To visualize the trade-off between precision and recall for different threshold t, it is possible to plot the two measurements

in Precision-Recall (PR) curves, where on the x-axis commonly the recall is plotted and on the y-axis the precision (an example is shown in Chapter 5, Figure 5.3). The area under the PR-curve (AUCPR) is used to compare different approach with each other, where approaches with higher AUCPR are assumed to be more accurate. To evaluate the trade-off between recall and precision the *F1-score* (Saito and Rehmsmeier, 2015) can be computed:

$$\text{F1-score} = 2 \cdot \frac{P \cdot R}{P + R}. \tag{2.3}$$

To find the optimal parameter setting based on the trade-off between recall and precision one usually took the maximal F1-score over all considered thresholds t.

Mean squared error

The *mean squared error* (MSE) is commonly used to measure how well a distribution fits to an observed outcome (James *et al.*, 2013, p. 29,30). MSE is defined as:

$$\text{MSE} = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2, \tag{2.4}$$

where $\hat{f}(x_i)$ is the prediction of the observation x_i, and y_i is the expected outcome. The smaller the MSE the better the predicted values fits to the expected ones.

Random variables

A *probability space* $(\Omega, \mathcal{A}, P)$ consists of the *sample space* Ω, which is described as the set of all possible outcomes (Deisenroth *et al.*, 2020, p. 174-175). For example, flipping a coin has the sample space $\Omega = \{head, tail\}$. An *event* is a subset of the sample space Ω and the *event space* $\mathcal{A}$ is given by all possible events (K. Blitzstein and Hwang, 2015, p. 3). Sticking to the flipping coin example, the event space is given as $\mathcal{A} = \{\{\}, \{head\}, \{tail\}, \{head, tail\}\}$. P is the probability that is assigned to each observable event $a \in \mathcal{A}$ and lies between $[0, 1]$. So for the flipping coin example, the probabilities for the events are given as following: $P(\{\}) = 0$, $P(\{tail\}) = 0.5$, $P(\{head\}) = 0.5$ and $P(\{head, tail\}) = 1$. In general, the probability of all possible outcomes $\omega \in \Omega$ is summing up to 1.
The probability space can be utilized to model some real world problems. To do so, we introduce *random variables*: Given the sample space Ω of an experiment, a random variable X is defined as a function from the sample space Ω to the real number $\mathbb{R}$ (K. Blitzstein and Hwang, 2015, p. 92), thus $X : \Omega \to \mathbb{R}$ (Deisen-

roth *et al.*, 2020, p. 175). For the flipping coin example, we can for instance introduce a random variable that counts how often *tail* is observed, hence the random variable X represents the two outcomes $X(tail) = 1$ and $X(head) = 0$. We are interested in the probability to observe a specific outcome (Deisenroth *et al.*, 2020, p. 175).
Within this thesis, we use random variables which are Laplace distributed (see Chapter 5), thus the sample space is continuous. A continuous random variable X can be identified by a function $f : \mathbb{R} \to \mathbb{R}^+$ if for any numbers a and b with $a \leq b$:

$$P(a \leq X \leq b) = \int_a^b f(x)dx. \tag{2.5}$$

Thus, $\int_{\infty}^{-\infty} f(x)\,dx = 1$. The function f is denoted as the *probability density function* (PDF) of X (Dekking, 2005, p. 57).
Further, the *cumulative distribution function* (CDF) of a random variable X (Dekking, 2005, p. 44) is given by the function $F : \mathbb{R} \to [0, 1]$, defined as

$$F(a) = P(X \leq a). \tag{2.6}$$

The derivation of the CDF equals the PDF.

The Laplace distribution

Our statistical approach in Chapter 5 is mainly based on the *Laplace distribution*, which is also known as double-exponential distribution. In the *Statistical Distributions*, Chapter 26, the properties of the Laplace distribution are summarized (Forbes *et al.*, 2010). The Laplace distribution $L(a, b)$ is depending on two parameters: the location a and the scale parameter b. The probability density function (PDF) of $L(a, b)$ is visualized in Figure 2.14 and given as:

$$f(x) = \frac{1}{2b} \cdot e^{\frac{-|x-a|}{b}} \tag{2.7}$$

and the corresponding cumulative distribution function is:

$$F(x) = \begin{cases} \frac{1}{2} \cdot e^{\frac{-|x-a|}{b}} & \text{if } x < a \\ 1 - \frac{1}{2} \cdot e^{\frac{-|x-a|}{b}} & \text{if } x \geq a \end{cases} \tag{2.8}$$

The Laplace distribution has a mean μ of a and the variance σ^2 is $2b^2$, thus $b = \sqrt{\frac{\sigma^2}{2}}$.

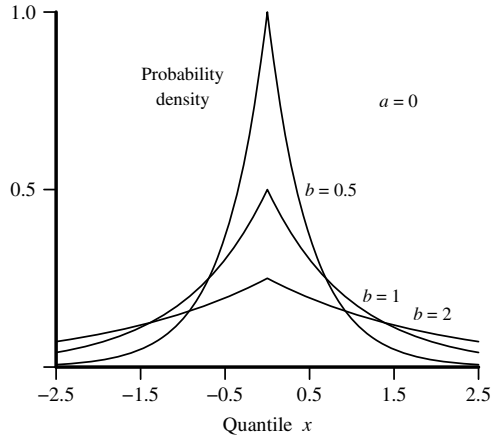


Figure 2.14: The probability density function of the Laplace distribution. A few examples of the PDF for varying b parameters are shown. The parameter a is set to zero. Figure taken from Forbes *et al.* (2010), Figure 26.

Maximum log-likelihood estimator

Given observed data that follows a known distribution with unknown parameters, the aim of a maximum log-likelihood estimator (MLE) is to choose this parameter such that the probability of observing the data is maximized (Deisenroth *et al.*, 2020, p. 265-268). We used a MLE in Chapter 5 to find the parameter b of our modified Laplace $L(n, b)$ distribution which fits the observed data for a TF model best. In general, the idea behind a MLE is to derive a function depending on the parameter θ, which is the parameter we are interested in, such that we can find the model that best fits to our data The observed data $x = \{x_1, x_2, ..., x_n\}$ can be modeled as random variable with a PDF $p(x|\theta)$ depending on θ. The likelihood function is the probability to observe x given the parameter θ:

$$L'(x|\theta) = p(x|\theta). \tag{2.9}$$

We assume that the observed data points are independent and identically distributed. The independence allows us to rewrite the likelihood as a product:

$$p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta). \tag{2.10}$$

Identically distributed means that each observed data point x_i is assumed to follow the same distribution with the same parameter θ. Commonly the es-

timation problem is represented as log-likelihood, since it is easier to solve, resulting in:

$$L(x|\theta) = log(p(x|\theta)) = log(\prod_{i=1}^{n} p(x_i|\theta)) = \sum_{i=1}^{n} log(p(x_i|\theta)). \tag{2.11}$$

The MLE can be obtained by finding the zeros of the derivative of $L(x|\theta)$ (Dekking, 2005, p. 316-318).

Hypothesis testing

Hypothesis testing is the statistical problem, whether a proposed hypothesis usually called null hypothesis (H_0), given the observed data is true or not (Lehmann and Romano, 2005; McDonald, 2014). If it is true, we want to accept it and reject the hypothesis otherwise. The opposite of H_0 is denoted as alternative hypothesis H_1. Various statistical test exists for hypothesis testing, the appropriate one is chosen depending on the kind of question one wants to answer, the data and the underlying distribution (Nayak and Hazra, 2011). Independent of the test used one can either make a correct decision or one of the two following errors: If H_0 is rejected but true, basically a false negative (FN) an error of first kind arises and if H_0 is accepted but false, so a false positive (FP) an error of second kind is made. The resulting consequences of these errors are not the same. Assuming, a new blood test was developed to check for the absence or presence of skin cancer. If an error of first kind occurs, the patient is predicted to have the disease, while she is actually healthy. However, a dermatologist will examine her skin and figure out that the prediction was wrong. In contrast, if an error of second kind occurs, a person suffering from skin cancer is predicted as healthy and no investigation or treatment is initiated.
Before applying the test usually a cutoff is set denoting the probability to incorrectly reject H_0. This cutoff is denoted as *significance level*. Usually the significance level is choose to be 0.05 or 0.01, meaning that with a chance of 5% or 1% the null hypothesis is rejected incorrectly. If the significance level is increased the chance of a false positive is increased. If we stick with the example from above, we would send more persons suffering from skin cancer home than with a smaller significance level. Simultaneously the chance of a false negative is decreased, thus in our example we would examine less health patients. In contrast if the significance level is decreased, we reduce the number of false positives, but make it also more difficult to detect a significant different from H_0.
A p-value is the resulting value of a hypothesis test, which gives the probability that we observe the data given H_0. If the p-value is less or equal the chosen

significance level, H_0 is rejected because it is assumed that it is unlikely to obtain the observed data under H_0.
To test if two independent samples which are assumed to follow the normal distribution have a similar mean, a unpaired two sided t-test is commonly applied (Zar, 2010, p. 130ff). A two-sided t-test for two random variables X and Y with the same number of observed samples is defined by:

$$t = \frac{\overline{X} - \overline{Y}}{S_p \cdot \sqrt{\frac{2}{n}}}, \tag{2.12}$$

where $\overline{X}$ and $\overline{Y}$ are the mean of the considered sample data for X and Y. The pooled standard error S_p describes the difference in the standard error between the mean of the two variables and is given as:

$$S_p = \sqrt{\frac{s_X^2 + s_Y^2}{2}}, \tag{2.13}$$

where s_X^2 and s_y^2 is the variance of the samples. To p-value for the resulting t can be looked up in a t-distribution table.
To evaluate whether two independent samples are identically distributed even if they do not follow the normal distribution a *Wilcoxon rank sum test* can be applied. Given two random variables X and Y with continuous CDF F and G. The Wilcoxon rank sum test, aims to test the null hypothesis H_0 that $F = G$. The alternative H_1 is that X is stochastically smaller than Y, defined as $F(a) > G(a)$ for all possible a (Mann and Whitney, 1947). The samples $x_1, ...x_n$ of X and $y_1, ..., y_m$ of Y are ranked. The idea of the Wilcoxon rank sum test (U) is to count how often $X > Y$, thus it can be defined as:

$$U_y = m \cdot n + \frac{m \cdot (m+1)}{2} - T_y, \tag{2.14}$$

where T_y is the sum of the ranks of the values of Y in an ordered sequence of X and Y. Since the labeling of X and Y is random, the test can also be computed as:

$$U_x = m \cdot n + \frac{n \cdot (n+1)}{2} - T_x, \tag{2.15}$$

where T_x is similar to T_Y but for the sum of the ranks of X. The minimum of U_y and U_x is the result of the Wilcoxon rank sum test and a p-value can be derived based on the distribution of U. For small samples the p-value can be looked up in so called recurrence tables based on the distribution of U, and for larger samples U is approximately normal distributed (Zar, 2010, p. 163-166).

Multiple testing correction

We often compute thousands of statistical tests, as for instance when computing a differential expression analysis for all known genes of the genome. As a logical consequence some will have a p-value smaller than the significance level just by chance even if all H_0's are actually true. There are two tests, which are commonly used to correct for this effect, the Bonferroni correction (McDonald, 2014, 257-259) which controls the type I error rate and the Benjamini-Hochberg procedure that handles the false discovery rate. The latter is often denoted as FDR correction.
The Bonferroni correction just decreases the significance level of each test to reduce the type I error rate (false positive rate). Therefore, the adjusted significance level is given by dividing the originally assumed significance level by the number of applied tests. Only p-values less or equal the adjusted significance level are assumed as significant. Thus, the correction is extremely strict when thousands of tests are performed.
A widely used alternative is the FDR correction (Benjamini and Hochberg, 1995). In a first step, the p-values of the n simultaneously performed tests are sorted in ascending order, such that $p_1 \leq p_2 \leq p_3 \leq .... \leq p_n$. The corresponding hypothesis of p_j with $j \in \{1, ..., n\}$ is H^i. The FDR-correction works as following: Find $k = \arg\max_i \ p_i \leq \frac{i}{n} \cdot \alpha$. Reject all H^i with $i \in \{1, ..., k\}$.
The FDR adjusted p-value is usually computes as adj-$p_i = min(p_i \cdot \frac{n}{i}, \text{adj-}p_{i+1})$, where p_{i+1} is the adjusted p-value of the next higher ranked p-value (McDonald, 2014, p. 260). The hypothesis of the FDR adjusted p-values are reject if adj-$p_i \leq \alpha$.

Monte-Carlo sampling

Monte-Carlo sampling (Harrison *et al.*, 2010; Kroese *et al.*, 2014) is a technique to estimate the likelihood to obtain an observation from an unknown distribution. The technique is applied to generate random simulations or processes *in silicio*. Often they are used to solve deterministic problems by generating random objects. The aim is to repeat the random process hundreds or thousand times to reliably describe its behavior in real life. Currently Monte-Carlo techniques are popular, since they are easy to apply, feasible on most machines and they can be used in various kinds of settings. For instance, an early application of tthe technique was to model the neutron transport process (Metropolis *et al.*, 1953) or to simulate chemical kinetics (Gillespie, 1977) but they are also used in the process of developing new materials, such as solar cells (Stenzel *et al.*, 2012) or lithium-ion batteries (Thiedmann *et al.*, 2011). Many more examples are summarized in Kroese *et al.* (2014).

Fisher's method

Several approaches exist to combine p-value from independent hypothesis tests with the same H_0 in a meta analysis. Among them is Fisher's method, which can be describe by a random variable X as following:

$$X(p_i, ..., p_m) = -2 \cdot \sum_{i=1}^{m} log(p_i), \tag{2.16}$$

where p_i denotes the p-value of the ith method (Heard and Rubin-Delanchy, 2018). Fisher's method follows the chi distribution, which is usually denoted as X^2, with $2m$ degree of freedom (Fisher, 1934). Thus, one can obtain the corresponding p-value by computing $1 - F_{2m}(x)$ with $F_{2m}(x)$ being the cumulative distribution function of the X^2 distribution.

2.2.2 Bioinformatic strategies to analyze sequencing data

In this section we outline how TF and histone ChIP-seq, RNA-seq and open chromatin data are commonly analyzed. For each kind of data several different tools exists, all having their advantages and drawbacks.

Histone and TF ChIP-seq

Independent of the kind of ChIP-seq experiment the sequenced reads from the studied TF or histone as well as the input are mapped to the reference genome using aligners, such as the burrows-wheeler alignment tool (BWA) (Li and Durbin, 2009) or bowtie2 (Langmead and Salzberg, 2012). The commonly used bowtie2 aligner consists of four steps. First, substrings and their complement of the reads are extracted and aligned to the genome using a FM-index (Ferragina and Manzini, 2001) based on a Burrows-Wheeler transformation (Burrows and Wheeler, 1994). Thereafter, the mapping positions need to be extracted from the index and finally the substrings are extended to the original reads using a dynamic programming approach.
The aim of a ChIP-seq experiment is to identify regions with an enriched signal of the TF or histone of interest in comparison to the input. The expected peak width depends on which kind of protein, TF or histone is chipped. While TFs tend to produce highly localized signals, usually denoted as narrow peaks, broad peaks are observed for some histone marks *e.g.* H3K36me3. Further, there are chipped targets, such as polymerase II which can have narrow as well as broad peaks (Furey, 2012). Therefore, different peak callers are developed, where some of them focus only on a specific type of peaks, such as SICER2 (Xu *et al.*, 2014), ZINBA (Rashid *et al.*, 2011), BroadPeak (Wang *et al.*, 2013) or Peakzilla (Bardet *et al.*, 2013). Additionally, there are peak callers that provide

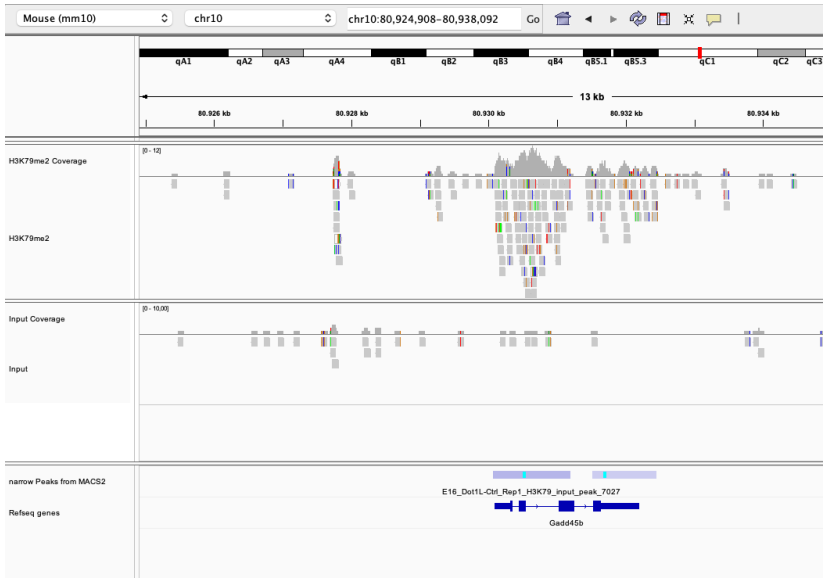


Figure 2.15: Visualizing the result of a histone ChIP-seq experiment. Using IGV the genomic region around the gene Gadd45b is deciphered. In the first row the location within the genome is outlined. For a H3K79me2 ChIP-seq experiment the resulting reads are aligned to the mouse reference genome using bowtie2 (second row). In comparison the input is shown (third row). Additionally, the peaks called with MACS2 are deciphered (fifth row, first track). The second track in the fifth row shows the annotated mouse genes from the reference genome. Two peaks within the genebody of Gadd45b are observed. The visualized data is taken from Cattaneo *et al.* (2022).

different options to handle peaks of different shapes, such as F-seq (Boyle *et al.*, 2008) or MACS2, which is an updated version of MACS (Zhang *et al.*, 2008). The underlying algorithm of MACS and MACS2 is based on the idea to correct for the strand asymmetry which is observable for ChIP-seq data around binding sites. This results in two slightly shifted peaks, the distance between the peaks is denoted as shift size. After shifting the reads from both strands towards each other according to their estimated shift size, it is assumed that the distribution of the reads can be model by a Poisson distribution. With a sliding window approach they detect windows with significant enrichment according to the Poisson distribution. The identified windows are merged to peaks and the input is used to estimate a false discovery rate for each peak. An example output of a histone ChIP-seq experiment visualized in the integrative genome viewer (IGV) is shown in Figure 2.15.

RNA-seq

There exist three different approaches to map the reads to the genome and determine the gene expression: (1) The genome mapping, (2) the transcriptome mapping and (3) the reference-free, *de novo* assembly approach, which are summarized in Figure 2.16.

Genome mapping: This approach can be used if an annotated genome of the species of interest is available. The mapping of RNA-reads to the genome is not as straightforward as for reads from DNA-sequencing, since reads can map across splice junctions (Kukurba and Montgomery, 2015). Thus, splicing-aware read aligners, such as STAR (Dobin *et al.*, 2012) or TopHat (Trapnell *et al.*, 2009) are used to map the reads to the genome. STAR, for instance, is based on a two step procedure. First, a seed search in uncompressed suffix arrays is performed, where STAR detects the longest sequence of a read that matches the reference genome. For the unmapped portion of the read the search is repeated. In the second step the separated reads are stitched together and scored.
After aligning the reads, the gene expression can be determined by counting the mapped reads. This can be done with a tools, such as Cufflinks (Trapnell *et al.*, 2012) either by providing an annotation file or without an annotation file, which allows the detection of novel transcripts.
Transcriptome mapping: If one is not interested in identifying novel transcripts, it is possible to directly map the reads to the transcriptome of the species of interest. Therefore, an aligner, such as bowtie2 can be utilized and the transcript identification and quantification can be done for instance with sailfish (Anders *et al.*, 2014) or kallisto (Bray *et al.*, 2016). Further, the method salmon (Patro *et al.*, 2017) became popular, which can be either used to align the reads and to quantify the transcript and gene expression or it can be applied to only quantify the transript and gene expression based on a precomputed alignment. Salmon applies a quasi-mapping, which efficiently and accurately approximate the alignment and correct for several biases, such as sequence- and position specific as well as fragment-GC content biases.
Reference-free/ de novo assembly: If no reference genome of the species of interest is available, it is necessary to first assemble the reads into transcripts. To quantify the expression of the different transcripts, one can use the previously outline strategy of transcriptome mapping. Throughout the thesis only human data is used, thus *de novo* assembly was not necessary and we mention the approach only for completeness.

After the reads are mapped to the genome and counted, the raw read counts need to be normalized to be comparable between expression levels among samples, since the count is dependent on transcript length and sequencing

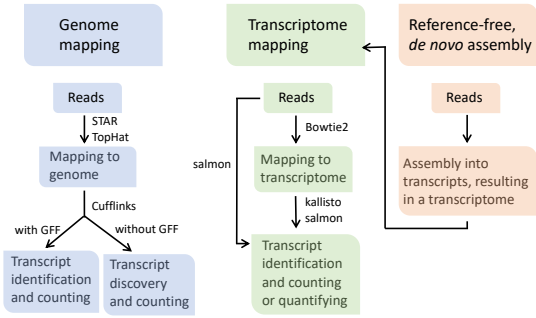


Figure 2.16: Computational approaches to determine gene expression. (blue) Genome mapping and (green) transcriptome mapping can be used if a reference genome/ transcriptome of the species of interest is available. (red) Reference-free, *de novo* assembly is applied if no reference genome of the studied species is known. Based on Figure 2 of Conesa *et al.* (2016), modified.

depth. Commonly used normalizations are reads per kilobase per million reads (RPKM) and transcripts per million (TPM). Since it has been shown, that RPKM is inconsistent among samples (Wagner *et al.*, 2012), we prefer to use TPM throughout this thesis. TPM is computed for each gene g as following:

$$TPM(g) = \frac{r_g \cdot rl \cdot 10^6}{gl_g \cdot T}, \tag{2.17}$$

where r_g is the number of reads mapped to the gene g, rl gives the read length in bp, and gl_g is the genes length in bp. T is the total number of transcripts sampled in a sequence run and defined as:

$$T = \sum_{g \in G} \frac{r_g \cdot rl}{gl_g}, \tag{2.18}$$

where G is the set of all considered genes.
Often one is interested in comparing the gene expression between different experimental conditions, such as cells treated with a different medium than the control cells (Tombor *et al.*, 2021), or studying the effect of a knockout/ silencing of a gene in comparison to the control (Cattaneo *et al.*, 2022). The aim is to identify genes expressions which significantly differ when comparing the two conditions. A possible approach to identify these genes which are differentially expressed is DESeq2 (Love *et al.*, 2014). The read count of each gene is assumed to follow a negative binomial distribution with mean μ and dispersion α, which are estimated based on the input data. For each gene a

shrinkage estimation is used to fit the dispersion and the log fold change to the predicted negative binomial distribution. A Wald test using the adapted log-fold change and the standard error is computed to determine the significant differentially expressed genes (DEG). As result the FDR corrected p-values are provided.

Open chromatin data

The first step of all open chromatin assays is to align the sequenced reads to the genome, which is usually done with bowtie2 (Langmead and Salzberg, 2012). Since MNase-seq only indirectly measures open chromatin, special software, such as DANPOS (Chen *et al.*, 2012) is developed to identify the accessible regions of the genome. For DNase1-seq usually algorithms are applied which have been developed to analyze ChIP-seq data, but without giving an input, such as MACS (Zhang *et al.*, 2008) or MACS2, F-seq (Boyle *et al.*, 2008) or HotSpot (Baek *et al.*, 2011). See also the review of Koohy *et al.* (2014), where they analyzed the performance of these peak callers for DNase1-seq data. JAMM (Ibrahim *et al.*, 2014) is a peak caller which is often used with DNA-seq or ATAC-seq. FAIRE-seq is mostly analyzed with MACS2. More detailed information can be found in the review of Tsompana and Buck (2014).

2.2.3 In-silico prediction of transcription factor binding

TFs are able to bind the DNA by recognizing short sequence patterns, denoted as TF motifs. These patterns can be described *in vitro* using high throughput methods like protein-binding microarrays (PBM) (Berger *et al.*, 2006) or SELEX (Jolma *et al.*, 2010), or *in vivo* using ChIP-based techniques (Lambert *et al.*, 2018; He *et al.*, 2015) (see Section 2.1.7). The identified TF binding preferences are summarized in a TF model, most prominent are Position Weight Matrices (PWMs) that for each position of the motif holds the weighted observed base count (Stormo, 2000). PWMs can be visualized as sequence logos (see Figure 2.17 and Section 2.2.3). However, there are other more complex TF models utilizing a bayesian network or other types of Markov models (reviewed in Boeva (2016)). Other proposed models are for instance the binding energy model (BEM) (Zhao *et al.*, 2012) or the transcription factor flexible model (TFFM) (Mathelier and Wasserman, 2013). Further there is the SLIM model (Keilwagen and Grau, 2015), which additionally provides a graphical visualization of the TF binding preferences and methods based on deep convolutional neural networks like DeepBind (Alipanahi *et al.*, 2015) or BPNet (Avsec *et al.*, 2021a).
Through out this thesis, we used PWMs to describe the binding preferences of TFs. They are widely used, available for hundreds of TFs for different species

and easy to apply. In Chapter 7 more reasons are outlined and discussed. In the following, we explain the details of how PWMs are constructed, outline the principle to compute the TF binding score and mention public resources storing TFs with known binding sites. Next, we introduce the technique, which we used in this thesis to derive a p-value for the TF binding score. Finally, we summarize two widely used methods to compute TF binding, which we also utilized in a few applications in this thesis.

Position weight matrices are used to describe TF binding characteristics

A position weight matrix (PWM) as well as the position frequency matrix (PFM) and position specific probability matrix (PPM), where the last two are preliminary stages of a PWM, are based on a set of aligned sequences S over the alphabet $\Sigma = \{A, C, T, G\}$. The sequences s_i with $i \in \{1, ..., n\}$ of S, which have the length m are assumed to be bound by a TF and thus represent the motif of a TF.
In a first step, the count of each base for each position in S are stored in a PFM. More formally, a PFM M which describes a TF motif of length m is a $4 \times m$ matrix, where each entry is defined as (Stormo, 2000):

$$M_{\sigma,j} = \sum_{k=1}^{n} I(X_{k,j} = \sigma), \tag{2.19}$$

where σ denotes a letter from the alphabet Σ, $j \in \{1, ..., m\}$, $X_{k,j}$ is the base of sequence s_k at position j and the indicator function $I(x = \sigma) = 1$ if $x = \sigma$ and 0 otherwise.
A PPM P describes the probability to observe a base at a given position in the TF motif, thus it can be derived from M where each entry of P is given by (Wasserman and Sandelin, 2004):

$$P_{\sigma,j} = \frac{M_{\sigma,j} + \epsilon}{\sum_{p \in \Sigma} M_{\sigma,j} + \epsilon}. \tag{2.20}$$

Usually a small pseudocount ϵ is added to each entry to avoid undefined behavior in case of zero entries.
Finally, P can be transformed to a PWM W, which is also known as position specific scoring matrix. Therefore, the background probability of the individual bases b_σ is taken into account by computing a log-ratio between the entries of P and the base frequency. Thus, an entry of W is defined as (Wasserman and Sandelin, 2004):

$$W_{\sigma,j} = log_2(\frac{P_{\sigma,j}}{b_\sigma}). \tag{2.21}$$

The TF binding score describes how well a sequence fits to a TF motif

Given a pattern L of length m over the alphabet Σ. To evaluate how likely it is that the pattern is bound by the TF according to a PWM W, one computes the *TF binding score* s as (Wasserman and Sandelin, 2004):

$$a = \sum_{j=1}^{m} W_{\sigma_{L,j},j}, \tag{2.22}$$

where $\sigma_{L,j}$ is the base of the pattern L at position j.
PWMs can be visualized in sequence logo introduced from Schneider and Stephens (1990), which gives an intuition whether a pattern may fit the motif of a TF or not. The idea is to illustrate each position of the motif by stacking the bases on top of each other. The more likely it is to observe a base at this position the higher the base is deciphered. The height of each base is depending to its frequency. The concept of the Shannon entropy H is utilized to measure the uncertainty at each position j of the motif represented as PPM (Schneider and Stephens, 1990):

$$H_j = -\sum_{\sigma \in \Sigma} P_{\sigma,j} * log_2(P_{\sigma,j}) \tag{2.23}$$

The information content R at position j is calculated as (Schneider and Stephens, 1990):

$$R_j = log_2(|\Sigma|) - H_j + e_n, \tag{2.24}$$

where e_n is a correcting factor necessary when only a few sequences are observed in the alignment. Finally, the height h of each letter σ at position j can be derived as (Schneider and Stephens, 1990)

$$h_{\sigma,j} = P_{\sigma,j} \cdot R_j. \tag{2.25}$$

An example of a resulting sequence logo is shown in Figure 2.17. Sequence logos can be created for protein sequences as well. The information content is measured in bits in the range from 0 to 2 in the case of DNA. When all bases occur with the same probability, than the information content is 0 bits, in contrast if only one base is observed at a position, the information content is 2 bits. Thus, the higher the information content, the more important the base is at a specific position.

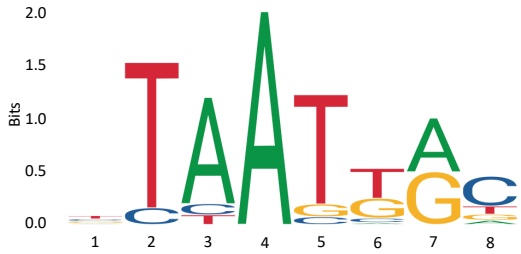


Figure 2.17: Sequence logo of the TF HOXB6. A graphical visualization of the motif of TF HOXB6 is shown, downloaded from the JASPAR database (Fornes *et al.*, 2019). The x-axis gives the position of the motif, while the y-axis shows the information content in bits.

PFMs/PWMs are available for hundreds of TFs

Publicly available databases, such as JASPAR (Castro-Mondragon *et al.*, 2021), Hocomoco (Kulakovskiy *et al.*, 2017) or Kellis (Kheradpour and Kellis, 2013) store hundreds of TF binding sites mostly represented as PFMs. The most successful one might be JASPAR, which is still under maintenance and regularly publishes new releases. For instance, the latest release of the JASPAR database (9th release) contains 841 non-redundant TF-motif pairs for vertebrates (Castro-Mondragon *et al.*, 2021). It is assumed that the human genome encodes for around 1500 TFs (Ignatieva *et al.*, 2015), thus not for all of them a motif is known yet. Within this thesis we either use the motifs from the JASPAR database or a collection of non-redundant motifs gathered from the JASPAR, Kellis and Hocomoco database.

Computation of a p-value for a TF binding score

Throughout this thesis, we utilize the dynamic programming approach from Beckstette *et al.* (2006) to calculate an exact TF binding score distribution such that we can derive a p-value for the TF binding score. It is necessary to compute a p-value for the TF binding score, since the score distribution of TFs differ from each other and thus is nit directly comparable. Possible reasons are the varying lengths and the different entropies of the TF motifs.
For a p-value π and a TF motif M of length m, we aim to identify the minimum given TF binding score threshold $T_{min}(\pi, M)$ such that all the larger scores than T describe the probability that M matches a random sequence of length m is at most π. To determine this threshold $T_{min}(\pi, M)$ for a given π Beckstette *et al.* (2006) utilized a dynamic programming approach that avoids to compute the complete probability distribution.

It is assumed that the bases $\sigma \in \Sigma$ in a sequence S of length n occur independent of each other with the probability $f(\sigma) = \frac{1}{n} \cdot |\{i \in [0, n-1] | S[i] = \sigma\}|$. We usually use the whole genome sequence to determine the frequency of the bases. The probability to observe a specific subsequence s of S of length m with $m << n$ can be computed as $P(s) = \prod_{j=0}^{m-1} f(s[j])$. The probability that a sequence s reaches a TF binding score t for a given TF motif M, is the sum of all sequences that achieve this score and can be expressed as:

$$P[sc(s, M) = t] = \sum_{s \in \Sigma^m : sc(s,M)=t} \prod_{j=0}^{m-1} f(s[j]), \tag{2.26}$$

where $sc(s, M)$ denotes the TF binding score between s and a given TF motif M and Σ^m are all possible words of length m over the alphabet Σ. The minimal threshold $T_{min}(\pi, M)$ for a given π can be described as:

$$\begin{aligned} T_{min}(\pi, M) = min\{t | sc_{min}(M) \leq t \leq sc_{max}(M), \\ P[sc(s, M) \geq t] \leq \pi\}, \end{aligned} \tag{2.27}$$

where $sc_{min}(M)$ is the smallest possible score for a given TF model M and $sc_{max}(M)$ the largest.
Using a dynamic programming approach, it is possible to efficiently compute the probability distribution of all TF binding scores. We consider all prefix TF motifs of M denoted by M_i with $i \in \{1, ..., m\}$. Assuming the probability distribution for M_{i-1} is known, it is possible to derive the distribution for M_i utilizing the assumption that the bases within a sequence are independent of each other.
Within the dynamic programming approach we use a matrix D, where each row i represents the probabilities for M_i. More explicit, D is $(m+1) \times (l+1)$ matrix, where m is the length of the TF motif and l is the number of all possible thresholds one can observe. To denote the number of observable thresholds, it is necessary to round the numbers in the TF motif to a fixed number of decimal places. In the first step, we initialize the first row of the matrix, such that the probability of observing a threshold $t \geq 0 = 1$. One can write it precisely as:

$$D_0(t) := \begin{cases} 1, & \text{if } t = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{2.28}$$

Each prefix of M, which corresponds to a row in D, can be described as $M_i : [0, i] \times \Sigma \to \mathbb{N}$ which is given by $M_i(\sigma, j) = M(\sigma, j)$ with $j \in \{1, .., i\}$ and $\sigma \in \Sigma$. For each possible threshold, it is determined whether it can be observed with the current M_i. The threshold t is only observable for M_i when in the previous row D_{i-1} a sequence s exists with a threshold of $t - M_{\sigma,i}$ for any $\sigma \in \Sigma$. This

can be computed as following:

$$D_i(t) = \sum_{\sigma \in \Sigma} D_{i-1}(t - M_{\sigma,i}) \cdot f(\sigma). \tag{2.29}$$

In the last row of the matrix D the probability to observe a threshold t for a given sequence s is given, thus $P[sc(s, M) = t] = D_m(t)$. The p-value is derived by the cumulative probability. Therefore, one sums up D_m until the threshold t is reached. The p-values are stored in a vector Q, which has as many entries as there are possible thresholds.
D is filled recursively from the bottom to the top until the cumulative probability exceed the given p-value threshold. An example of this approach is explained in Figure 2.18.

Hit-based and affinity-based approaches to determine a TF binding score

In this section, we briefly outline two software tools commonly used to determine the TF binding score. However, the two approaches differ fundamentally. Find individual motif occurrences (FIMO) (Grant *et al.*, 2011) is a hit-based approach, meaning it determines whether a sequence is bound by a TF or not. In contrast, transcription factor affinity prediction (*TRAP*) (Roider *et al.*, 2007) computes the average binding probability of a TF for a sequence, and thus is a so-called affinity-based approach.

FIMO: Given a set of TF motifs and a set of sequences, the method determines all TF binding sites with a p-value ≤ 0.01 (default setting). FIMO computes the TF binding score as well as the p-value similarly as outlined in the previous section. A bootstrap method is used to determine the p-value's false discovery rate and correct it for multiple testing by providing a q-value. This value gives the minimal false discovery rate at which the p-value is assumed as significant. As a result, a list of TFs binding sites, $p-$ and q-values is provided.

***TRAP*:** The method provides an affinity score for each tested TF and sequence, giving the expected average number of bound TF molecules. Therefore, TRAP is based on a biophysical model assuming that the complex formation of a TF and a sequence S can be described in the equilibrium $TF + S \Leftrightarrow TF \cdot S$. The fraction of bound sites can be computed as

$$p(S) = \frac{[TF \cdot S]}{[S] + [TF \cdot S]} = \frac{K \cdot [TF]}{1 + K \cdot [TF]}, \tag{2.30}$$

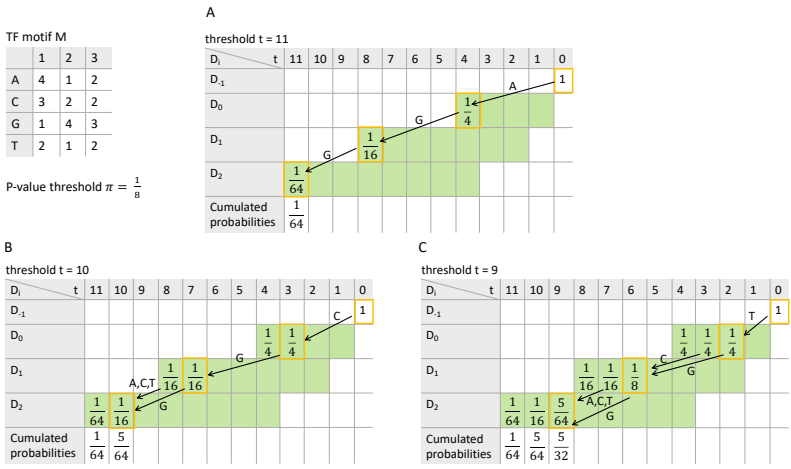


Figure 2.18: Detailed example for the approach from Beckstette *et al.* (2006). An example how to derive the score threshold $T_{min}(\pi, M)$ for a TF motif M of length 3 and a p-value threshold of $\pi = \frac{1}{8}$ is visualized. The green marked fields are those which one needs to compute to derive the full score distribution. We assume that all bases are observed with the same frequency, thus $f(\sigma) = \frac{1}{4}\ \forall\ \sigma \in \Sigma$. The cells that are computed within the current step are marked by an orange square. (**A**) In a first step, the probability to derive a score $t = 11$ is recursively computed. Only the sequence *AGG* achieves a score of 11 and is computed as following: $D_2(11) = D_1(8) \cdot f(G)$, $D_1(8) = D_0(4) \cdot f(G)$ and $D_0(4) = D_{-1}(0) \cdot f(A) = 1 \cdot \frac{1}{4}$. As a result $D_2 = \frac{1}{64}$, which equals the cumulated probabilities for $t = 11$. In (**B**) and (**C**) all sequences that achieve a score $t = 10$ and $t = 9$ are considered, respectively. The previously computed value for $D_1(8)$, $D_1(7)$ and $D_0(3)$ can be utilized without recomputation. (**C**) The cumulated probability sums up to $\frac{5}{32}$ and exceeds the given p-value threshold $\pi = \frac{1}{8}$, thus the algorithm stops. The score threshold for the given p-value $\pi = \frac{1}{8}$ is 10. Based on Figure 11 of Beckstette *et al.* (2006)

where K is the site specific equilibrium constant. The author derived how the parameters can be estimated using a probabilistic framework which is based on the idea of Berg and von Hippel (1987). The binding probability is determined for each site in a sequence and summed up to an overall affinity. The affinities can be directly compared between different TFs as well as between sequences of different length, since the affinities are length normalized.

2.2.4 TF motif enrichment analysis

A TF motif enrichment analysis aims to identify TFs that regulate the transcription of a set of genes. Therefore the enrichment of known TF motifs in

regulatory elements or promoter regions of the gene set is detected (McLeay and Bailey, 2010). The known TF motifs can be gathered from public TF motif databases, such as JASPAR or Hocomoco (see Section 2.2.3).
A typical task which is solved by a TF motif enrichment analysis is to identify TFs that regulate a set of co-regulated genes. Within this thesis, we applied a TF motif enrichment analysis to detect TFs that are involved in the regulatory changes of circPLOD2 depleted pericytes (details are outlined in Section 4.3). We used the regulatory elements of the DEGs between control pericytes and circPLOD2 depleted pericytes to predict enriched TFs.
A TF motif enrichment analysis can also be utilized to determine the unknown motif of a chipped TF. To do so, it is either assumed that the motif of the TF is similar to a known one, hence the known motifs are used within the TF enrichment analysis or *de novo* motif discovery tool are applied to determine a set of possible motifs. Then the TF motif enrichment is used to refine this set of *de novo* motifs (see also Section 3.1).
In the following two widely used TF motif enrichment tools are outlined.

CentriMo - Centrality of Motifs

CentriMo is a method to link motifs to TFs especially for TF ChIP-seq data, assuming that the binding sites of the ChIP-ed TF occurs more enriched in the center of the peaks than in the flanking regions. As input, *CentriMo* expects a set of motifs and a set of sequences of equal length identified with a TF ChIP-seq experiment. In a first step, *CentriMo* scans for each motif the given sequences for the significant binding sites for all possible positions. For each sequence only the binding site with the highest binding score is considered. Assuming that the best binding site of a motif can occur at any position in the sequence, a binomial enrichment test is applied to derive a p-value for the central enrichment of a motif. Therefore, over all sequences it is counted how often the best binding site of a motif appeared within the central w positions in comparison to the flanking regions. Since it is not known what the best positions are to consider as central, different widths are tested. For each of them a p-value is computed and adapted for multiple testing using Bonferroni correction. The method returns, among other things, a ranked list holding a p-value representing the central enrichment for each tested TF-motif pair .

PASTAA - predicting associated transcription factors from annotated affinities

PASTAA is a method that incorporates the binding affinity of a motif with a biological signal, which can be various kind of biological information related to the considered genes, phenotype or utilized experiments. In their publica-

tion the authors used the ChIP-chip binding values as biological signal. In the circPLOD2 depleted pericyte example mentioned above we used the open chromatin signal of a REM in the studied cell type as signal value and in Chapter 3 we utilize the TF ChIP-seq signal value. As input the method expects a set of motifs and a set of sequences, where each sequence is associated to a biological signal. In a first step, the sequences are ranked according to their predicted binding affinity using *TRAP* (Roider *et al.*, 2007), which computes the average binding probability of a given motif for the entire sequence. The higher the binding affinity the more likely that the TF binds the sequence. Additionally, the same set of sequences is ranked according to the biological signal. The idea is to apply a cut-off to both lists and to check if there is a significant overlap between the considered sequences. Thus, a good candidate motif provides a high binding affinity to a set of sequences, which also have a high biological signal. The enrichment is computed for different cutoffs using a hypergeometric test to determine the significance between the sequences with high binding affinity and the sequences with high biological signal. As a result, the methods provide a p-value based on the hypergeometric test for each tested TF-motif pair.

2.2.5 Similarity measurement and clustering of TF binding motifs

In this thesis, we apply the similarity measurement and cluster algorithm for PFMs developed by Pape *et al.* (2008) in different chapters. For example, our domain score of MASSIF in Chapter 3 is based on the similarity measurement S^{max} defined in their work. Further, we clustered the TF motifs from the JASPAR database using their cluster algorithm.
Pape *et al.* (2008) developed a software package, called *Mosta*, which conducts motif similarity computation and motif clustering. Two PFMs X and Y are assumed to be similar if they describe a similar binding site, or to put it in another way, if they have a high number of overlapping hits in a random sequence. *Mostas'* similarity concept between two PFMs is based on an overlap probability $\gamma_{X,Y}(k)$ at position k of X and Y as well as on the probabilities of independent hits for X and Y at this position k, denoted by α_X and α_Y. The overlap probability $\gamma_{X,Y}(k)$ is the sum of the probabilities for all possible words $x \in X$ and $y \in Y$ to overlap at a position k. In addition, α_X is the probability that the words $x \in X$ occur in a background model, equally for Y. Applying the logarithm to the ratio of the overlap probability and the product of the probabilities of independent hits for X and Y, leads to the similarity $S_{X,Y}$:

$$S_{X,Y} = log\left(\frac{\gamma_{X,Y}(k)}{\alpha_X \cdot \alpha_Y}\right). \tag{2.31}$$

The ratio describes the probability to observe two hits assuming the motifs X and Y represent a similar binding site, normalized by the probability to observe hits for X and Y assuming they do not describe a similar binding site. In MASSIF domain score we applied a concept, also provided by *Mosta* called S^{max} which is based on Eq. 2.31. S^{max} is a maximization over all possible k's and it also considers the reverse complement of the motifs:

$$S^{max}(X,Y) = max(\max_k S_{X,Y}(k), \max_k S_{X',Y}(k), \max_k S_{Y,X}(k), \max_k S_{Y',X}(k)) \tag{2.32}$$

where X' and Y' are the reverse complement of X and Y, respectively.
Based on the definition of the similarity between two PFMs, *Mosta* determines clusters in a set of motifs. Each resulting cluster is a set of motifs which is represented by a consensus motif. Initially all considered motifs are interpreted as a separate cluster containing one motif which is also the consensus motif at the same time. Then in a greedy fashion, clusters are merged using S^{max} as a similarity measure. The procedure stops if the motifs in a cluster are not similar enough to the cluster consensus motif. The algorithm terminates if all pairs of consensus motifs are considered at least once or if the similarity of the remaining motif pairs is too low.

2.2.6 Identification of regulatory elements and their associated target genes

To understand genetic regulation it is crucial to identify REMs and to detect the genes they may regulate. Various kinds of experimental assays exist to predict REMs resulting in diverse sets of annotations, which are often publicly accessible. In Section 4.1.1 publicly available databases that hold REMs are summarized.
Often the problem is split in two tasks, where first the REMs are identified, which are then linked to target genes. When DNase1-seq, TF and histone ChIP-seq data is utilized to predict the REMs usually peak callers are involved to detect open chromatin regions or regions with an enriched TF or histone signal. Even though there are many peak callers specifically developed for the needs of the different experimental assays, they still have a few drawbacks. Often it is not trivial to decide which cutoff is most suitable to determine peaks in comparison to the input signal. Also, it is unclear at which level of enrichment a region is biologically relevant (Chen and Zhang, 2010). Further, cell cycle stage (Liu *et al.*, 2017) and cell numbers (Gilfillan *et al.*, 2012) can have an effect on the prediction of enriched sites.
To avoid all these drawbacks, a method called *STITCHIT* (Schmidt *et al.*, 2021) has been developed, which is not relying on a peak caller. Since we used

it in Chapter 4 as basis of our EpiRegio webserver, we provide more details about it in the following.

STITCHIT: A integrative approach to identify gene-specific REMs

STITCHIT is an algorithm to identify REMs and their potential target genes by analyzing epigenetic signals in relation to gene expression among several cell types or tissues. To do so, paired DNase1-seq and expression data of various samples are utilized. In more detail, the signal of the DNase1-seq data in a user-defined region around each gene is inspected to identify segments with a varying signal between the used samples. The aim is to choose the segments in such a way that the samples can be separated according to their target gene expression. Initially each segment is just one base pair long. Adjacent segments are combined, if the variance between the DNase1 signal of the segments is low with respect to the target gene expression. The minimal description length principle (Grünwald, 2007), an approach known from the information theory, is used to assign a score to each segment and two evaluate the costs of combining to adjacent ones. To identify the optimal segmentation for each gene, a dynamic programming approach is used. The segments from the optimal segmentation with a high correlation between the DNase1 signal of the segment itself and the target gene expression are used as features for a two-level machine learning approach.
First a linear regression model using elastic net regulation is applied for feature selections. Thereby the DNase1-signal within the segments are used to predict the target gene expression. Based on a cross validation procedure, models where the segments are not able to capture the information to predict the target gene expression well enough are excluded. For the remaining models, only the segments with a non-zero regression coefficient are kept. We denoted these segments as regulatory elements (REMs). To not only evaluate how well the REMs explain the gene expression, but also to assess their individual contribution on the target gene expression and their significance, an Ordinary Least Square (OLS) model is trained on the selected REMs. The resulting regression coefficients and the corresponding p-values are reported, and can be used in downstream analyses.

2.2.7 Regulatory SNPs (rSNPs) and approaches to detect them

In the Chapters 5 and 6, we are interested in so called regulatory SNPs (rSNPs). These are non-coding SNPs that have an impact on the gene regulation by affecting one or more TF binding sites. They either can increase the binding strength of a TF to the DNA or even create new binding sites. In contrast, rSNPs also can decrease the binding strength or disrupt a binding site of a TF.

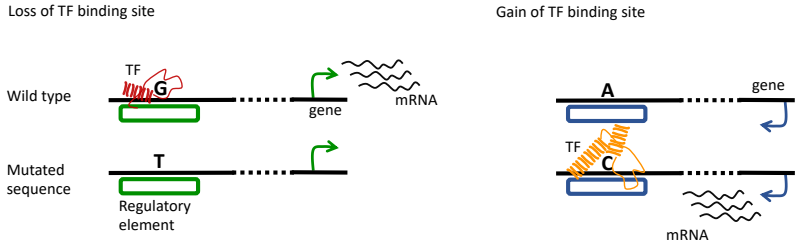


Figure 2.19: Examples of regulatory SNPs. In this figure two examples for regulatory SNPs (rSNPs) are outlined. (left) Caused by a rSNP a TF binding site is lost and the TF can not longer bind the sequence, which lead to an change in the target genes expression. (right) A gain of a TF binding site, induced by a rSNP that results in a enhanced gene expression, is visualized.

Two possible examples of a rSNP are visualized in Figure 2.19.
Several methods exist to evaluate the effect of a SNP on TF binding sites. We summarize them in Section 5.1. In this section we provide more details about the methods *sTRAP* (Manke *et al.*, 2010), *is-rSNP* (Macintyre *et al.*, 2010) and *atSNP* (Zuo *et al.*, 2015) that are closely related to our statistical approach in Chapter 5.
The three approaches only rely on the DNA sequence and a TF motif, which is represented by a PWM. The idea is to compare the binding strength between the wildtype sequence S^1 containing the wildtype allele and the mutated sequence S^2 holding the alternative allele.

sTRAP: An approach to rank affinity changes

In the work of Manke *et al.* (2010), they use their previously proposed approach *TRAP* (see Section 2.2.3) to compute the binding affinity of a PWM to a sequence. Given a TF model M and the sequences S^1 and S^2 of the same length than the TF model, a differential TF binding score is derived as following:

$$r_M = log_{10}(\frac{p_M(S^1)}{p_M(S^2)}), \tag{2.33}$$

where $p_M(S^i)$ denotes the p-value of the binding affinity of a sequence S^i, with $i \in \{1, 2\}$. High positive values pinpoint to an increase in the binding affinity induced by the SNP, whereas highly negative values are associated with a decrease of the binding affinity of the studied TF. The score is computed for all subsequences overlapping the SNP and the one with the highest absolute score is kept. For each considered TF motif the analysis is repeated. To assess

which TF is most affected by the studied SNP, the resulting maximal scores are ranked.

is-rSNP: Computation of statistical significant differential TF binding scores

The general idea of this approach is similar to the *sTRAP*. However, is-rSNPs uses a hit-based approach to compute the TF binding score and it provides a significance for the derived differential TF binding score.
Given a TF motif M of length m and sequences S^1 and S^2 of length $2m-1$. In a first step *is-rSNP* determines the position in S^1 and S^2, where the TF motif M achieves the highest TF binding score. To do so, they use a sliding window approach to compute the TF binding score for the TF model M and the corresponding p-value of all subsequences. For the window which has the maximal TF binding score that is also significant (p-value ≤ 0.001) either in wildtype or the mutated sequence, a differential TF binding score is derived as log-ratio similar to the one in Manke *et al.* (2010). To derive a p-value for the resulting score based on the given TF motif M, the complete differential TF binding score distribution is determined. Therefore, the differential TF binding score is computed for all possible sequences of the length m, where for each sequence all possible SNPs are considered. Based on the resulting distribution, a p-value for an observed differential TF binding score can be determined. Since an exact p-value distribution is assessed, this approach results in an algorithm with quadratic runtime in m.

atSNP: Affinity testing for regulatory SNPs

The aim of *atSNP* is to overcome the slow computation of a p-value for the differential TF binding score. Furthermore, the details how to compute the differential TF binding score differ slightly from the approach of *is-rSNP*.
Given a TF motif M of length m and sequences S^1 and S^2 of length $2m-1$, *atSNP* computes for each sequence the maximal TF binding score, including the reverse complement. To derive a p-value for the maximal TF binding scores of S^1 and S^2, it is assumed that the subsequences, which overlapping the SNP can be described by a first order Markov model, utilizing the 30 bp up and downstream of the SNP position. Thus, they do not assume that the positions within the sequence occur independent of each other. To evaluate the TF binding score distribution they use importance sampling instead of the classical Monte-Carlo sampling. They are mainly interested in the p-values of high TF binding scores, which one observes rarely by random sampling, thus a large number of simulations would be necessary to derive a reliable TF binding score distribution with a Monte-Carlo sampling (Chan *et al.*, 2010). Importance

sampling is an alternative where the focus is more on the events of interest. Thereby, the number of simulations is reduced, which results in a significant speed up. The differential TF binding score is computed as the log-ratio of the p-values of the maximal TF binding scores of the wildtype and the alternative sequence. A p-value for the differential TF binding score is determined using importance sampling to estimate the distribution of the differential TF binding scores.

2.2.8 How SNPs are linked to genes

Non-coding SNPs can have an impact on gene regulation (see Section 2.1.9), but it is not trivial to associate these SNPs to candidate genes. GWAS result in a summary statistic, providing the information how likely it is associated to the studied phenotype for each tested SNP. To identify which genes might be affected, gene-based association tests can be applied. The idea of such a test is to detect genes to which multiple SNP are linked to. These SNPs are often only associated with a medium strength to the studied phenotype and thus not prioritized by a GWAS (Li *et al.*, 2011). However, complex diseases are often caused by the interplay of many common variants with rather small effects (Eichler *et al.*, 2010). In a gene-based test, usually common SNPs and their SNPs in LD are analyzed using different statistical approaches (Li *et al.*, 2011; Liu *et al.*, 2010). For each gene the SNPs within a specific window are included in the analysis, for instance Li *et al.* (2011) used a 5 kb window. The result is a list of genes ranked according to their predicted impact in the studied phenotype.
An alternative approach is to evaluate the significant SNPs from the GWAS summary result and their SNPs in LD individually. Before linking SNPs to genes one can determine which of the SNPs might have a regulatory effect. A strategy is to utilize cell type or tissue specific eQTLs *e.g* downloaded from GTEx (GTEx Consortium, 2017). This is often done as downstream analysis of an GWAS using for instance a colocalization analysis (Hormozdiari *et al.*, 2016; Chen *et al.*, 2017). Another possibility to do so, is to use chromatin accessibility quantitative trait loci (caQTLs) which can be detected in ATAC-seq analyses. If an open chromatin region contains a SNP, one observes two kinds of reads, one holding the wildtype allele and one holding the alternative allele. Based on the varying read counts for the alleles of the SNP it is possible to derive information about the SNPs functionality on the open chromatin status (Broekema *et al.*, 2020; Kumasaka *et al.*, 2015). Many caQTLs detected in primary CD4+ T-cell show a strong enrichment to candidate SNPs associated to the autoimmune system (Qu *et al.*, 2015a). Further, it is possible to detect if a SNP is regulatory by predicting if a TF binding site is affected, using for instance atSNP, which is explained in the previous Section 2.2.7 or our statis-

tical approach introduced in Chapter 5.
Several strategies to associate a given set of SNPs to proximal and distal target genes candidates are known and commonly used (Dey *et al.*, 2022; Gazal *et al.*, 2022). Among them the less specific window-based strategy, where within a used defined region all SNPs are linked to a gene of interest (Zhu and Stephens, 2018; Kim *et al.*, 2019). Often SNPs in the promoter region of a gene are assumed to affect this gene. Furthermore, SNPs can be linked to genes based on previously defined enhancers, which are associated to target genes using various kind of cell type or tissue specificity epigenetic data. A few resources that hold enhancers associated to target genes are outlined in Chapter 4. Based on the cell type or tissue of interest there might be a specific dataset available, such as the cis-regulatory atlas of the mouse immune system (Yoshida *et al.*, 2019). Additionally, enhancer gene links can be derived using the activity-by-contact (ABC) score (Fulco *et al.*, 2019) or its generalized version (Hecker *et al.*, 2023) which uses open chromatin and/or histone ChIP-seq data combined with Hi-C data.

2.2.9 DisGeNET database, a resource for disease genomics

The DisGeNET database (Pinero *et al.*, 2015; Piñero *et al.*, 2019) is a collection of disease and variant associated genes, which are collected from various sources, such as UniProt (The UniProt Consortium, 2017), Orphanet (Rath *et al.*, 2012), ClinGen (Rehm *et al.*, 2015), ClinVar (Landrum *et al.*, 2017), GWAS Catalog (Buniello *et al.*, 2018), GWAS DB (Li *et al.*, 2015) as well as scientific literature and data gathered from animal models. The database is highly transparent, meaning the origin of the associations is given by providing the original database the data was retrieved from, the number of publications and their NCBI PMIDs as well as a text field summarizing the evidence. In total the current release (v7.0) contains $1,134,942$ disease gene association for $21,671$ genes and $30,170$ diseases and traits, and $369,554$ variant disease associations for $194,515$ variants and $14,155$ diseases and traits (https://www.disgenet.org). For each association a score is given which is based on the evidence which supports it. The authors of DisGeNET provide an R package (https://www.disgenet.org/static/disgenet2r/disgenet2r.html) that allows to explore the database, among other applications a disease enrichment analysis is provided. Given a list of genes, the disease enrichment test detects which genes are enriched in diseases or traits in comparison to all annotated associations. To do so, Fisher's exact test is used for each trait or disease within the database, and the resulting p-values are FDR corrected using the Benjamini-Hochberg method.

Chapter 3

Incorporating DNA binding domain information to TF motif enrichment analysis

In this chapter we introduce our method to improve TF motif enrichment analysis including DNA-binding domain information called MASSIF. Parts of the text and the figures are taken from our paper *Improved linking of motifs to their TFs using domain information* (Baumgarten *et al.*, 2019) published in *Bioinformatics*. We also provide the source code and a detailed description of how to use it in our GitHub repository https://github.com/SchulzLab/MASSIF. The contribution of all authors is given below:

- Conduction of the computational analyses, implementation of MASSIF, visualizing of the result, and writing the text: myself
- Discussion and design of the approach, evaluation of the results: Marcel H. Schulz (M.H.S), Florian Schmidt (F.S) and myself
- Proofreading and adjustment of the text: F.S and M.H.S

3.1 Common strategies to infer TF motifs in comparison to our approach

TFs are able to bind to the DNA by recognizing specific patterns with their tertiary protein structures denoted as DNA-binding domains (DBDs). As outlined in detail in Section 2.1.7, TFs are involved in gene regulation by binding to regulatory elements like promoters or enhancers. Often the first step to un-

derstand the regulatory effect of a TF on a gene, a phenotype or a disease, is to computationally localize the TF's binding site within the genome. To do so, the binding motif of the TF of interest must be known.
The binding sites of a TF can be experimentally determined by performing a TF ChIP-seq experiment. To pinpoint the motif of the chipped TF, *de novo* motif discovery tools are commonly used to identify short sequence patterns within the peak regions resulting from the ChIP-seq experiment (reviewed in (Tompa *et al.*, 2005) or (Tran and Huang, 2014)). Usually the result of a *de novo* motif discovery tool is a list of motifs that are significantly enriched in these sequences. Among the detected motifs can be not only the true motif, but also motifs of co-factors of the TF of interest, or repetitive sequences.
Alternatively, methods like *Clover* (Chen *et al.*, 2004), *PASTAA* (Roider *et al.*, 2009), *CentriMo* (Bailey and Machanick, 2012),*i-CisTarget* (Potier *et al.*, 2015), *REGGEA* (Kehl *et al.*, 2018) or *iRegulon* (Janky *et al.*, 2014) make use of the increasing number of already known motifs linked to TFs to detect enriched motifs in the given sequences. The known motifs are usually taken from motif databases like JASPAR (Khan *et al.*, 2017), TRANSFAC (Matys *et al.*, 2006) or Hocomoco (Kulakovskiy *et al.*, 2017) (see Section 2.2.3). Some of these methods can also handle *de novo* motifs as input, thus they are used as a follow-up analysis to refine the motifs resulting from a *de novo* motif discovery tool.
Additionally, there is a closely related field of research that investigates the prediction of the motif of a TF independently of any associated DNA-sequence. To the best of our knowledge one of the first studies that linked TFs to their motifs only relies on information derived from the amino acid sequence of the TFs, and thus also incorporates the DBD (Tan *et al.*, 2005). The approach is motivated by the observation that TFs associated to the same DBD are in general more similar to each other in terms of their amino acid sequence and therefore tend to bind to similar motifs. Tan *et al.* (2005) used a probabilistic framework including DBD similarity of TFs and comparative information for prediction in *E.coli*. A study which is based on a support vector regression model showed that for some DBDs useful features from the protein sequence can be derived to predict the motif of a TF (Schröder *et al.*, 2010). However, a study by Zamanighomi *et al.* (2017) showed that using only features derived from the DBDs yields a high number of false positives. To overcome this problem, they combined the DBD-based information with a probabilistic model of motifs hits using epigenetic data.
To summarize, utilizing information derived from the DBD of a TF of interest was successfully used to infer the corresponding motif often independently of any TF-associated sequences. However, we noticed that commonly used tools that associate motifs to TFs based on TF-associated sequences are not using the powerful DBD information.

3.2 Massif - motif association with domain information

In this section, we introduce a new method called Massif - motif association with domain information - that incorporates the DBD information to existing tools that link motifs to TFs with the aim to improve their prediction.
In the following, we first outline Massif's statistical approach, which is comparable to the statistic used by Tan *et al.* (2005). Thereby, we explain how a score is derived to evaluate how likely the predicted TF-motif pair is the correct one. Further, we show that well-known and commonly used tools combined with Massif's statistic based on DBD information increase their performance.

3.2.1 Overview of Massif

Our method Massif aims to improve the prediction of TF-motif pairs by combining DBD information with tools that link motifs to TFs. We observed that TFs with the same DBD often have similar motifs, for instance the HOX-factors, which are extremely similar to each other, all belong to the *Homeo domain factors*. Lambert *et al.* (2018) visualized the similarity of known TF motifs within a heatmap (see Figure 2A in their review), which showed a high similarity between the motifs of the TFs with the same DBD. We hypothesize that a candidate motif linked to a TF of interest is more likely to be the correct one, if it is similar to already known motifs associated to TFs with the same DBD as the TF of interest. Based on this idea, we developed a statistic. We constructed a DBD collection that holds TF-motif pairs sorted according to their known DBD. We used the collection to obtain motifs linked to TFs with the same DBD as the TF of interest. Based on this information we can compute a similarity measure between the studied TF and the linked candidate motif that evaluates how well the motif fits to the TF's 3-dimensional characteristics given by the DNA-binding domain. We call this similarity measure *domain score*, which can also be represented as a p-value.
Our method expects as input (1) a list of candidate motifs, which can be either *de novo* motifs or motifs from a motif database (2) the DBD of the studied TF, (3) TF-associated regions e.g. ChIP-seq data and (4) a biological signal per region of interest *e.g.* the signal value provided by ChIP-seq data. The domain score can be used in two different ways: (1) Using Fisher's method as a meta analysis (Fisher, 1934) (see Section 2.2.1) to combine the domain score with the p-value of predicted TF-motif pairs of existing tools. (2) Applying the domain score as a filter to reduce the list of candidate motifs before applying an existing tool. An overview is shown in Figure 3.1.

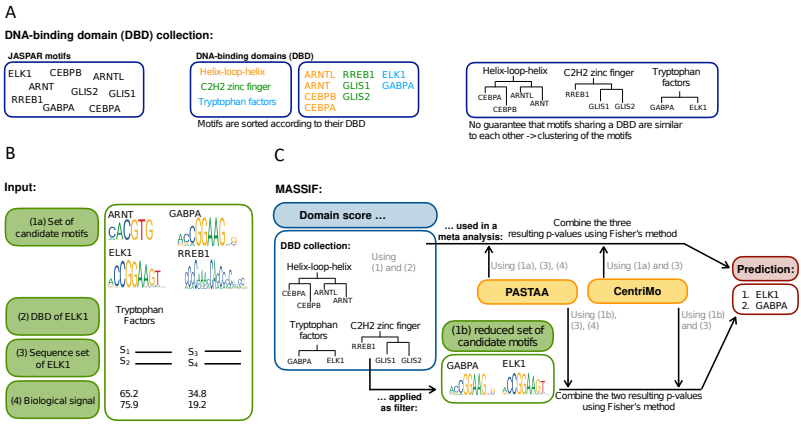


Figure 3.1: Overview of the DBD collection and our method Massif. (**A**) The steps how to build the DBD collection are outlined. (**B**) The required input to our method is shown. (**C**) The workflow of MASSIF is illustrated for the setup where the domain score is used in a meta analysis (upper part) and for the setup where the domain score is applied as a filter (lower part).

3.2.2 Preparation of the DNA-binding domain collection

We build the DBD collection using the TF-motif pairs from the JASPAR database. Each TF of the database is associated to a TF class and a TF family according to the TFClass system (Wingender *et al.*, 2013) (see also Section 2.1.7). We decided to use the TF class information as proxy for the DBD, since it might be easier to provide the TF class of the studied TF than the more specific TF family. The first step to construct the DBD collection is to separate the JASPAR TFs and hence indirectly the motifs, according to their DBDs. The 515 TF-motif pairs from the JASPAR database belong to 30 different DBDs (listed in the Appendix Table 1), where between 1 and 140 TF-motif pairs are assigned to the same DBD.

We noticed some variance between the motifs of TFs assigned to the same class of DBD, especially if many TFs are assigned to it. It is important to represent the diversity of motifs for a DBD as accurately as possible to provide reliable domain scores. On the other hand, we want to combine highly similar motifs to remove redundancy in the motifs associated to a DBD. Therefore, we cluster the motifs assigned to the same DBD into distinct groups using the cluster algorithm introduced by Pape *et al.* (2008) (see Section 2.2.5). For each DBD we get a set of clusters D_l, with $l \in \{1, ..., 30\}$ where 30 is the number of observed DBDs. Then $\mathcal{D}$ is defined as the set of all DBDs $\mathcal{D} = \{D_1, ..., D_{30}\}$. We can identify each DBD by their their corresponding set of clusters and vice

versa. Each set of clusters D_l consist of a varying number n_l of clusters c_l^k with $k \in \{1, ..., n_l\}$ depending on how similar the motifs within the DBD are. Thus, we can write $D_l = \{c_l^1, ..., c_l^{n_l}\}$. In addition, the motifs within a cluster c_l^k are represented by a consensus motif $M_{c_l^k}$.
The DBD collection allows us to look up a set of consensus motifs, which are commonly observed for TFs of a specific DNA-binding domain.

3.2.3 Definition of the domain score

To calculate the domain score between a candidate motif and the studied input TF, we need to be aware of the DBD of this TF. In general, for most of the TFs the DBD is known and can be looked up, for instance in UniProt (The UniProt Consortium, 2017). Otherwise, if the DNA- or protein sequence is known, the DBD can be predicted using tools like SMART (Letunic *et al.*, 2015) or UniPROBE (Hume *et al.*, 2015).
Let's consider the DBD D_l of the studied TF, then we can use the DBD collection to look up the consensus motifs of the set of clusters $M_{c_l^k}, k \in \{1, ..., n_l\}$ associated to D_l. The domain score provides for a candidate motif a score indicating how similar this motif is to the most similar consensus motif of the DBD of the current TF. In more detail, we calculate the similarities S^{max} between the PFM P of the candidate motif and all consensus motifs $M_{c_l^k}$ of the DBD of the current TF. S^{max} describes the similarity between two PFMs and was introduced by Pape *et al.* (2008). We provided more details about this similarity in Section 2.2.5. So, the set of similarities between a motif described as a PFM M and the consensus motifs of D_l can be computed as:

$$sim(M, D_l) = \left\{ S^{max}(M, M_{c_l^k}) \mid c_l^k \in D_l \right\}. \tag{3.1}$$

Among all calculated similarities, we pick the highest one. Since the similarity value also depends on the motif itself, we divide the maximal similarity by the sum of all maximal similarities over all DBDs for the current motif P. Thus, we get the following equation for the domain score I_D:

$$I_D(M, D_l) = \frac{\max sim(M, D_l)}{\sum_{m=1}^{|D|} \max sim(M, D_m)}. \tag{3.2}$$

We assume that the higher the similarity, the more likely it is that the candidate motif fits to the current TF. Figure 3.2 illustrates an example how to calculate the domain score. To enable a better interpretation of the domain scores and to allow us to use them in a statistical test, the domain scores are represented as p-values π:

$$\pi := P(I_D(M, D_l) \geq x | H_0), \tag{3.3}$$

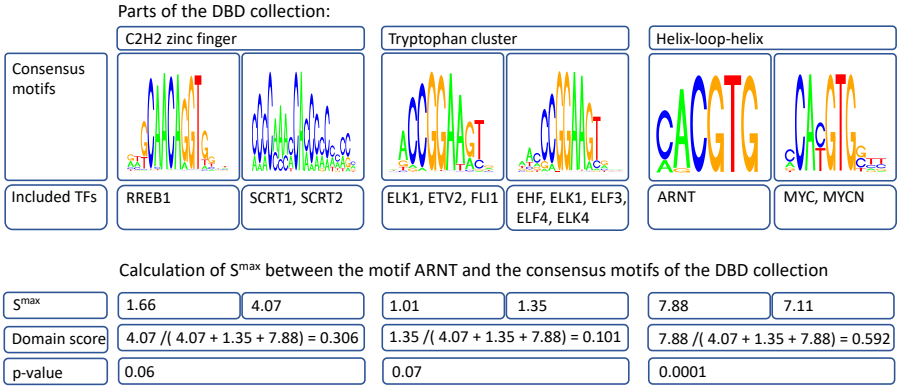


Figure 3.2: Calculation of the DBD score. Example how to calculate the domain score on a small part of the DBD collection. (Top) For each DBD the corresponding consensus motifs are shown and the TFs within this cluster are listed below. (Bottom) The similarities S^{max} between the PFM of the TF ARNT and the consensus motifs of all DBDs are computed. The calculated domain scores and the corresponding p-values assuming the candidate TF has the DBD 'Tryptophan cluster', 'C2H2 zinc finger' or 'Helix-loop-helix' are shown.

where x is an observed value of the domain score and the null hypothesis H_0 is defined as: *the candidate motif linked to the studied TF does not fit well to motifs of TFs associated to the same DBD as the studied TF.*
Since we have no analytical description of H_0, we approximate it by using Monte Carlo sampling (see Section 2.2.1). We randomly sampled 100,000 PFMs with a length between 6 and 21. The length of a motif was drawn from a normal distribution, since the motif length of the JASPAR motifs followed this distribution. For each entry of the matrix representing the motif, we picked a number based on a uniform distribution between 1 and 100, and divided each column by the column sum, so that we get PFMs. By doing so, we generated a set of randomly sampled PFMs R.
For each random motif $r \in R$ we calculated the domain score x for a given DBD D_l and estimated a p-value for this score as:

$$p(x, D_l) = \frac{|\{r | I_D(r, D_l) \geq x, r \in R\}|}{|R|}. \tag{3.4}$$

Basically, we count how often a random motif has an observed domain score for a given DBD that is higher or equal than the score x and divide it by the total number of motifs in R.

The domain score I_D can either be used in a meta analysis or as a motif filter. In the following both setups are explained in more detail.

Domain scores used in a meta analysis

Tools that link motifs to TFs relying on TF-associated sequences usually return a list of motifs describing the likelihood that the motifs are over-represented in the set of input sequences. For each motif, a p-value is given that is used to compare the motifs between each other and to also rank them. For each motif within this list, we determined the domain score I_D (Equation (3.2)) and the corresponding p-value as explained in Equation (3.4). A meta analysis is performed by using Fisher's method (see Section 2.2.1) to combine the p-values of the used tools and the domain score:

$$X(p_1, ..., p_m) = -2 \cdot \sum_{i=1}^{m} log(p_i), \tag{3.5}$$

where p_i represents the p-value of the ith method as well as the p-value of the domain score. Thus, $m - 1$ methods that link TFs to motifs are used. Fisher's method follows the χ^2 distribution with $2k$ degrees of freedom (Fisher, 1934). As a consequence, we can obtain the corresponding p-value by computing $1 - F_{2k}(x)$, where $F_{2k}(x)$ is the cumulative distribution function of the χ^2 distribution.

Domain score used as a motif filter

As an alternative, one can also apply the domain score in a pre-processing manner to reduce the number of motifs used in the tools linking motifs to TFs. Given the DBD of the TF of interest, we can use the domain score to exclude motifs which are unlikely to be observed for TFs with this DBD. By doing so, we can increase the reliability of the predicted TF-motif pairs and reduce the number of false positives. We chose a threshold t for the domain score p-value, so that we can decide for which of the input motifs the domain score is significant ($\pi \leq t$). We excluded all motifs from the motif set which are not significant. Finally, we applied the considered tools on the reduced motif set and, when we used more than one tool, we combined the resulting p-values with Fisher's method.

	C	P	CP	MCP	M_C	M_P	M_{CP}
Meta analysis			✓	✓			✓
Domain score as filter					✓	✓	✓

Table 3.1: Explanation and shortcuts of the different setup. 'C' and 'P' denote 'CentriMo' and 'PASTAA', respectively. In the setup 'CP', the results of *CentriMo* and *PASTAA* are combined within a meta analysis. 'M' followed by capital letter(s) denotes the use of the domain score in a meta analysis and 'M' followed by subscript letter(s) the use of the domain score as filter.

3.3 Evaluation of existing TF motif enrichment tools combined with MASSIF's domain score

In the next section, we evaluate the performance of existing TF-motif tools incorporating MASSIF's domain score on TF ChIP-seq data with motifs gathered from a TF motif database and *de novo* learned motifs.

3.3.1 Using state of the art tools to link motifs to TFs

Our proposed method MASSIF is only able to improve the performance of existing TF-motif tools, that associate motifs to TFs based on TF-associated sequences. Among these tools *CentriMo* (Bailey and Machanick, 2012) and *PASTAA* (Roider *et al.*, 2009) are the most widely used ones (McLeay and Bailey, 2010), thus, we decided to use them also for our evaluation. However, our method is not specific to these two tools and can also be combined with other tools that link motifs to TFs. In Section 2.2.4 a detailed description of *CentriMo* and *PASTAA* is provided. In brief, *CentriMo* is designed to analyze ChIP-seq data and prioritizes motifs found in the middle of peak regions utilizing a binomial test that compares motif occurrence in the center and the flanking regions. *PASTAA* is a rank based tool, that does not only include TF-associated sequences, but also biological information to identify TF-motif pairs. In addition, these tools are user-friendly, easy to apply and their runtime is tolerable for large sequence sets.

3.3.2 Combining MASSIF's domain score with existing TF motif enrichment tools improves their performances

To evaluate our method and to compare it to the performance of *CentriMo* and *PASTAA*, we used 102 TF ChIP-seq datasets for which a linked motif for each chipped TF is known (see also Appendix Section A.1.2). Thus, for a predicted TF-motif pair we can evaluate if it is the correct one. To control if MASSIF improves the results of the considered tools, we test multiple variations, for

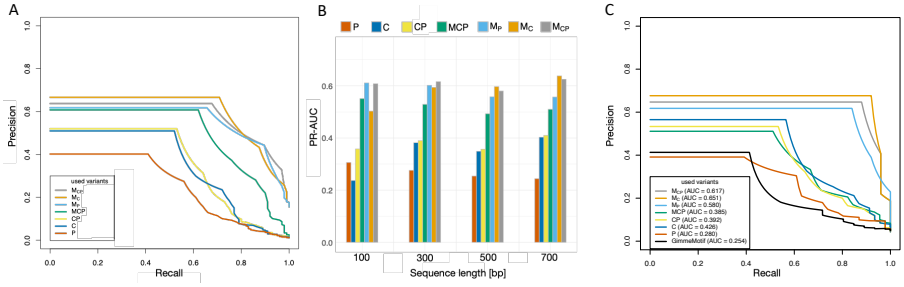


Figure 3.3: Results of the ENCODE dataset for different setup. (**A**) and (**B**) show the results of MASSIF using Hocomoco motifs for 102 ENCODE ChIP-seq dataset. (**A**) Precision-Recall curves that visualizes the performance to correctly link motifs to TFs for sequence length 500 bp. (**B**) Bar plot illustrating the AUCPR (y-axis) of the results for different sequence lengths (x-axis). (**C**) Precision-Recall curves of the results of MASSIF with *de novo* identified motifs for 46 ENCODE ChIP-seq datasets (sequence length 300 bp).

which we developed the naming scheme shown in Table 3.1.
We first tested MASSIF with the known motifs of the Hocomoc database and then in a more difficult scenario, where we use *de novo* motifs as input motifs.

Results using Hocomoco motifs as input

We ran MASSIF for all setups shown in Table 3.1 on four sequence sets differing in length (100 bp, 300 bp 500 bp and 700 bp) and used the motifs of the Hocomoco database as input. To assess the performance of the tested variations, we determined for how many TFs a motif was linked correctly. To account for similarity between different Hocomoco motifs in the evaluation, we combined similar motifs by clustering (Pape *et al.*, 2008). Linked motifs that belong to the same cluster as the true motif were counted as correctly linked. The results are shown in Figure 3.3, where Figure 3.3A shows a AUCPR curve for all used setups for sequence length 500 bp and Figure 3.3B the PR-AUC for all setups for the different sequence lengths as bar plot.
To get a first clue how well *PASTAA* and *CentriMo* perform, we ran them without any modifications. We noticed that the AUCPR of *PASTAA* are highest (≈ 0.310) for the shortest sequence length and dropped for the longer ones. For *CentriMo* we found the opposite effect. The lowest AUCPR (0.237) was achieved for a sequence length of 100 bp, and the performance was improved for longer sequences *e.g.* a AUCPR of 0.382 was achieved for a sequence length of 300 bp.
Next, we combined the predictions for a TF-motif pair of *PASTAA* and *CentriMo* in a meta analysis using Fisher's method, denoted as *CP*. The AUCPR of

CP was similar compared to the AUCPR of *CentriMo* (except sequence length 100 bp), whereas an improvement of the AUCPR of *PASTAA* was observed. In general, the performance of *CP* was less susceptible to varying sequence length compared to the individual results of *CentriMo* or *PASTAA*.
To test if we can further improve the AUCPR of *CP*, we added the p-values of the domain scores to the meta analysis. We refer to this setup as *MCP*, which resulted in a clear improvement of the AUCPR compared to *CP*. Especially for the sequence length 100 bp the increase of correctly linked motifs was substantial. The average improvement of the AUCPR of *MCP* over *CP* was around 0.142 for all sequence lengths.
Next, we tested how the *domain score* used as a filter affects the prediction. Therefore, we evaluated the results of *PASTAA* and *CentriMo* on the reduced motif set, where all motifs with domain score p-value > 0.001 were excluded. The setups are termed M_P and M_C, respectively. Interestingly, these setups improve the performance of many analysis. For sequence length 100 bp the AUCPR of M_P was 0.048 lower than the AUCPR of MCP. On average over all sequence lengths the AUCPR for M_C compared to *CentriMo* improved by around 0.204 and M_P compared to *PASTAA* by around 0.324. Still, both tools showed varying performance with varying sequence length. For instance, the AUCPR of M_P dropped by around 0.104 after increasing the sequence length from 100 bp to 700 bp.
Finally, we combined the setups M_P and M_C within the meta analysis, and refer to it as M_{CP}. Compared to M_C the performance was similar, except for the shortest sequence length, where an increase of the AUCPR of 0.105 was achieved. For M_P longer sequences resulted in the largest improvement. Additionally, differences in the AUCPR for varying sequence lengths were small compared to the meta analysis or *CentriMo* and *PASTAA*. We observed stable AUCPR over all sequences for M_{CP}, but for longer sequences the AUCPR of M_C was slightly higher.

We conclude that adding the domain score leads to a substantial improvement of linking motifs to TFs, where the filter-based method leads to a slightly higher improvement than the meta analysis.

Results using *de novo* motifs as input

As an alternative to the Hocomoco motifs, we investigated how *de novo* motifs affect the performance of Massif. To determine *de novo* motifs on the ENCODE ChIP-seq datasets, we applied a tool named *GimmeMotifs* (van Heeringen and Veenstra, 2010) on sequences of length 300 bp. We used this method for the following reasons: It identifies *de novo* motifs for ChIP-seq datasets in appropriate time and combines several *de novo* motif discovery tools within

one method (used tools: MDmodule, MEME, Weeder, MotifSampler, trawler, Improbizer, BioProspector, Posmo, ChIPMunk, AMD, Homer, XXmotif). Additionally, *GimmeMotifs* clusters the resulting motifs to decrease the number of redundant ones. The algorithm was able to identify *de novo* motifs for 46 out of the 102 downloaded ChIP-seq datasets. We applied MASSIF on these datasets using the identified *de novo* motifs as input.
To evaluate how accurately the different variations perform, we followed the same strategy as before, and assessed how many motifs were correctly linked to a TF. To determine which *de novo* motif is the correct one, we calculated the similarity for each of them to the known motif of the TF of the current ChIP-seq dataset. For this we used the similarity function *sstat* from Pape *et al.* (2008), which computes the similarity between two PFMs. The *de novo* motif that is most similar to the motif of the TF is assumed to be the correct one. Figure 3.3C shows the PR curve for the different setups. We added a black curve which represents the performance of *GimmeMotifs* by sorting the *de novo* motifs according to their internal enrichment value.
We observed that in general the tendency of the results of MASSIF using *de novo* motifs as input were similar to the results based on the known TF-motif pairs shown in Figure 3.3A and B. The setups which use the domain score as a filter again performed best. Also the AUCPR from *CentriMo*, *PASTAA* and *CP* were similar in comparison to MASSIF with Hocomoco motifs as input. Only the performance of *MCP* decreased.
In general, we noticed that some of the correctly linked *de novo* motifs were extremely similar to the true motif, an example is shown in Figure 3.4A. Further, we observed that the filtering was able to remove *de novo* motifs that were highly different from the true one (see Figure 3.4B). For 12 out of 46 cases the filtering rejected even all *de novo* motifs because of low similarity to any known motifs of the DBD of the current TF (p-value > 0.001).
To conclude, the filtering can be applied as a quality control to eliminate *de novo* motifs that are found to be enriched in the sequences, but most likely do not represent the true motif of the TF. Adding the domain score, especially as filter, leads to a considerable improvement of the performance of *CentriMo* and *PASTAA*.

3.3.3 The number of motifs within a DBD affects the domain score

In the following section, we investigate the domain score distribution for randomly sampled motifs and the wrongly predicted TF-motif pairs.

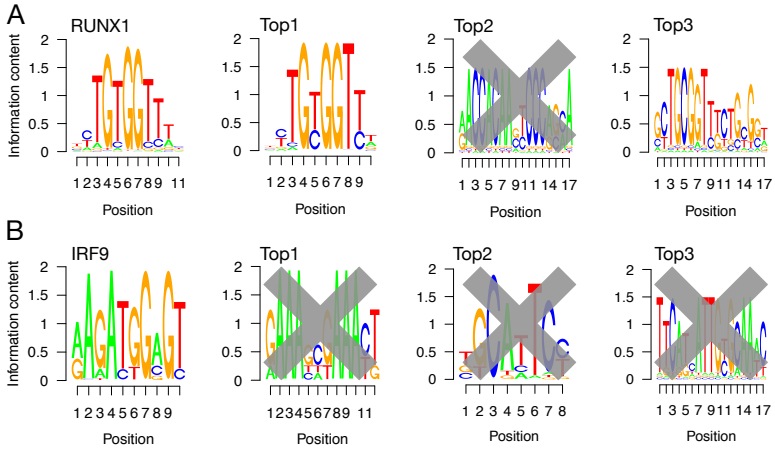


Figure 3.4: Examples of *de novo* identified motifs in comparison to the true ones. The first row visualizes the known motif of the current TF of interest, the following motifs are *de novo* motifs sorted according to their similarity to the true motif. Motifs that are eliminated throughout the filtering step are crossed out. (**A**) An example is shown where the *de novo* predicted motifs on the ENCODE ChIP-seq data for the TF *RUNX1* are similar to the known motif. The filtering excludes a *de novo* motif that does not fit properly to the true one. (**B**) An examples is outlined, where the filtering eliminates all *de novo* motifs. In such cases MASSIF is not able to make a prediction.

DBD with less motifs provides more reliable domain scores

We studied the domain score distribution of the randomly sampled motifs separately for each DBD and noticed that the distributions differ from each other. In Figure 3.5A the domain score distribution for the DBD of *C2H2 zinc finger factors* and *STAT domain factors* are shown as example. Generally, the domain scores of DBDs that contain fewer motifs have a smaller mean than DBDs with a large number of motifs. An explanation for this effect could be that the DBDs with a large number of motifs typically consist of several clusters. Thus, it is more likely that a random motif gets a higher domain score, just by chance, because of the large number of computed motif comparisons.

DBD with less motifs might better distinguish true TF-motif pair from incorrect one's

To get a better intuition of how the domain score improves the AUCPR of the used tools *CentriMo* and *PASTAA* we analyzed the motifs which are incorrectly linked to a TF using the motifs from the Hocomoco database. In particular,

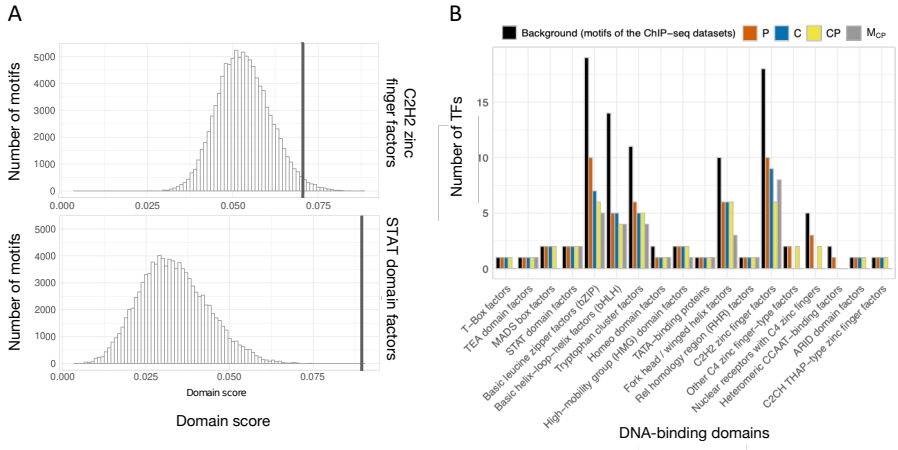


Figure 3.5: Domain score distribution and analysis of incorrectly predicted TF-motif pairs. (**A**) The distributions of the domain scores for two DBDs, namely 'C2H2 zinc finger factors' and 'STAT domain factors' are shown. The black vertical lines mark the smallest values of the domain score of all real motifs of the DBDs 'C2H2 zinc finger factors' and 'STAT domain factors'. (**B**) Analysis of the motifs incorrectly linked to a TF (y-axis) of some setups denoted by the colors for sequence length 500 bp. The background represents the number of TFs per DBD of the input ChIP-seq datasets. The plot separates TFs by their DBDs (x-axis).

we wanted to investigate for which DBDs using the domain score is helpful. In Figure 3.5B the number of motifs that are incorrectly linked to a TF per DBD are shown. The black bars illustrate the number of TFs of the input ChIP-seq dataset per DBD denoted as background. If we compare the number of motifs that are incorrectly linked to a TF of the different setups, we notice several interesting points.

First, in all cases *CP* is able to be at least as good as *CentriMo* or *PASTAA*. For four DBDs *Basic leucine zipper factors*, *Basic helix-loop-helix factors*, *Tryptophan cluster factors* and *C2H2 zinc finger factors* the performance of *CP* achieves a better result than *CentriMo* or *PASTAA* alone. Interestingly, in two cases, namely *Other C4 zinc-type factors* and *Nuclear receptors with C4 zinc finger*, the performance of *CP* is poorer compared to the best used tool, that links motifs to TFs.

The performance for M_{CP} is in most cases at least as good as the best considered setup or improves the performance. For the DBD *C2H2 zinc finger factors* M_{CP} links two motifs incorrectly, which are correctly linked from CP. In addition, we observe that M_{CP} leads to an improvement of 7 DBDs in such a way that all linked motifs are correctly associated, whereas without using the

domain score *PASTAA* or *CentriMo* link one motif incorrectly. Interestingly, five out of these seven DBDs contain only up to five motifs in the DBD collection, whereas the other two contain 11 and 36 motifs. These findings suggest that for DBDs, which contain less motifs, the domain score strongly influences performance, which is consistent with the observation of Figure 3.5A. There, we notice that DBDs containing less motifs, are able to better distinguish if a motif is correctly associated to a DBD or a false positive one.

We conclude that combining *CentriMo* and *PASTAA* in a meta analysis improves the result in comparison to the individual tools. Further, the reduced motif set which results from the filtering based on the domain score is much more specific than the original motif set. This leads to a more accurate prediction, especially for DBDs with less motifs.

3.4 Discussion

In this chapter, we analyzed how the DBD information of TFs can be used to improve the performance of existing approaches, that link motifs to TFs relying on TF-associated sequences. Our tool Massif is based on the idea that a correctly linked motif for a TF of interest is similar to motifs of TFs with the same DBD as the studied TF. We introduced Massif's *domain score*, which computes a similarity measure and evaluates how likely it is that a candidate motif fits to the studied TF. The domain score can either be used in a meta analysis or as a filter to reduced the set of input motifs. Our method Massif incorporates the tools *CentriMo* and *PASTAA* for which we showed that the domain score improves the AUCPR of these tools.
The domain score can not only be applied to *CentriMo* and *PASTAA*, but also to improve the performance of any tool that links motifs to TFs if it is based on a matrix representation like PFMs. The idea of the domain score is also suitable for more complex representations of DNA motifs if appropriate similarity measures exist. Even if a tool does not provide a p-value as part of its analysis, our domain score can be used as a filter to improve results.
Massif requires the DBD of the studied TF as additional information. As already mentioned before, it is practical to determine the DBD of a TF even if the DBD is not listed in a protein database like UniProt by predicting it using protein domain profiles (El-Gebali *et al.*, 2019). However, in cases where the DBD information cannot be derived or is not included in our current DBD collection, our method cannot be used.
We utilized 102 human sequence sets from ChIP-seq experiments to evaluate the performance of our method on a realistic setting. Within this work, we used as motif sets either the motifs available in the Hocomoco database or we

predicted them *de novo*. We avoided the motifs from the JASPAR database, since we build the DBD collection on it. However, in practice, any motif set can be used, from another motif database or experimentally derived or *de novo* motifs. Further, we investigated sequence sets with differing length, where our method produced stable results for all lengths, which is important, as peak length may vary with experiment quality and the used peak caller.

The advantage of using the domain score in a meta analysis is that no threshold has to be selected. However, using Fisher's method might not be the optimal statistical approach to combine the domain score with the p-values of tools that link motifs to TFs based on TF-associated sequences. The smallest possible p-value of the domain score using default parameters is 10^{-5}, since the corresponding distribution is based on $100,000$ random motifs. Compared to the smallest possible p-values of *CentriMo* and *PASTAA*, this p-value is rather big. Fisher's method is more sensitive to smaller p-values (Heard and Rubin-Delanchy, 2018). Thus, depending on the resolution, this kind of approach may favor p-values from either *CentriMo* or *PASTAA*. On the other side, if we use the domain score as a filter and combine the results of *PASTAA* and *CentriMo*, this bias is less problematic, as the p-values of both tools are in the same range. This could possibly explain why using the domain score as a filter achieves better AUCPR than using the domain score within Fisher's method. However, for the filter-based approach we need to choose a p-value threshold. Within this work we set it as 0.001.

The idea to combine two or more tools is not new. It has been applied in the context of *de novo* motif discovery tools or methods that link motifs to TFs based on TF-associated sequences, *e.g.*, MotifViz (Fu *et al.*, 2004), completeMOTIFs (Kuttippurathu *et al.*, 2010) or MEME-ChIP (Ma *et al.*, 2014). Nevertheless, none of these approaches combine the results of multiple tools in a statistical analysis.

Clearly, the results are depending on the tools used. CentriMo, for instance, uses flanking regions around the peaks to simulate a background distribution. Choosing too narrow peaks leads to worse results since the background is not represented reasonably, as we observed for the sequence length 100 bp and 300 bp. In contrast, *PASTAA* uses *TRAP* (Roider *et al.*, 2007) to estimate a TF binding affinity for all sites in each sequence. The longer the sequences, the more sites are incorporated, which may lead to a less accurate affinity computation. As we have shown in our evaluation, longer sequences decrease the number of correctly linked motifs.

Whether *de novo* or known motifs should be used, is not entirely clear, since both options have advantages and disadvantages. Using known motifs is fast and leads to more reliable results, however, no new motifs can be identified. Hence, it might be a problem if the motif of the studied TF is not similar to the set of known input motifs. As a solution, one can predict *de novo* motifs and

use them as input motif set. In our evaluation for roughly 50 of the datasets we were not able to identify significant *de novo* motifs. Further, the identified *de novo* motifs might include motifs of co-factors of the studied TF, motifs derived from repetitive sequences and, depending on the quality of the data and characteristics of the studied TF, it might not be possible to identify the true motif by the *de novo* discovery algorithm. Using the domain score as filter, we might be able to exclude most of the motifs identified by side effects. However, for one-third of the remaining datasets all predicted motifs were excluded by the filtering. Thus, with *de novo* predicted motifs only for a fraction of the 102 TF we managed to predict a candidate motif.
Refining the DBD collection might be a possibility to improve the predictions of our method. Large DBDs like *Homeo domain factors* or *C2H2 zinc finger factors* could be split into multiple smaller ones by using for instance the TF familiy information instead of the TF class. As outlined in the evaluation, we observed the tendency that DBDs with less motifs have the potential for a higher improvement than DBDs with more motifs (see Figure 3.5B). We decided against further refining the DBD collection, since it might become difficult for the user to provide the DBD of the studied TF.
To sum up, we presented that a commonly available and easy to access information of the TF, namely the DBD can be used as additional information to significantly improve the performance of tools that link motifs to TFs based on TF-associated sequences. The source code of method MASSIF is freely available online at https://github.com/SchulzLab/MASSIF.

Chapter 4

Webserver for the categorization of human regulatory elements

In this chapter the EpiRegio webserver, a resource of regulatory elements (REMs) associated to target genes is introduced. Further, two applications of EpiRegio are presented based on collaborations with the labs of Kathi Zarnack and Stefanie Dimmeler, as well as the lab of Manuel Kaulich.

4.1 EpiRegio: Analysis and retrieval of regulatory elements linked to genes

Our EpiRegio webserver is published as part of *Nucleic Acids Research*s' webserver issue with the title *EpiRegio: Analysis and retrieval of regulatory elements linked to genes*. Parts of the text and figures are similar as in our publication. Additional to the webserver (https://epiregio.de), we provide a detailed documentation available at: https://epiregiodb.readthedocs .io/en/latest/. The work was done in a close cooperation between Dennis Hecker (D.H), Sivarajan Karunanithi (S.K) and myself, thus we also share the first authorship of the publication. Below a list is provided, giving the contributions of all authors of our EpiRegio paper:

- REST API access, computing model score, DNase1 signal and cell type and tissue score, create mySQL database and import data: myself
- Parsing *STITCHIT* results in suitable *.csv* format: S.K and myself
- Documentation and figures: S.K, D.H and myself

- Define queries, decisions of the general functionality, defining the model and cell type or tissue score and testing the website: S. K, D.H, Marcel H. Schulz (M.H.S) and myself
- mySQL database schema: D.H, S.K, M.H.S, Markus List (M.L) and myself
- Data collection and *STITCHIT* analyses: Florian Schmidt (F.S)
- General advice regarding Django, mySQL and webserver setup: M.L
- Django web interface and query processing: D.H
- Setting up technical details to allow the public access of EpiRegio: S.K
- Writing the paper and proofreading: all authors

4.1.1 Related publicly available databases that contain regulatory elements

Regulatory elements (REMs) are non-coding genomic regions, like enhancers or promoters, that harbor TF binding sites and thereby regulate gene expression. Additionally, enhancers can be transcribed resulting in enhancer RNA (eRNA) (Mikhaylichenko *et al.*, 2018). More details about REMs are provided in Section 2.1.7.
The identification of REMs is not straightforward, and there is no method available yet that determines them with absolute certainty. Various kinds of epigenetic measurements are utilized in several combinations to predict REMs, leading to a diverse set of annotation approaches (Creyghton *et al.*, 2010; He *et al.*, 2010; Li *et al.*, 2012; Arnold *et al.*, 2013). Hence, there are also a number of publicly available websites providing REMs. For instance the *VISTA Enhancer Browser* contains experimentally validated cell type-specific REMs tested *in vivo* in transgenic mice. The investigated REMs are selected based on the conservation of the REM in other species or based on ChIP-seq data. The database is continuously extended, however, the number of interactions is still limited because of the time-consuming validation (Visel *et al.*, 2007).
A database containing REMs based on human data is the *FANTOM5 Human Enhancers* atlas (Andersson *et al.*, 2014), which detects cell type or tissue specific REMs based on Cap Analysis of Gene Expression (CAGE) data generated within the FANTOM5 project. Also, *HACER* (Wang *et al.*, 2018a) an atlas of human REMs incorporates CAGE data from FANTOM5, but also GRO/PRO-seq data, which is more sensitive to unstable eRNA than the CAGE technique. The presence of eRNAs is a clear indicator for enhancers, however, REMs that are not transcribed can not be detected with these techniques.
As a result, there are databases available that incorporate data from more than one experimental assay. An example is *READb* (Cai *et al.*, 2019) that holds REMs identified based on self-transcribing active regulatory region sequencing (STARR-seq) and massively parallel reporter assay (MPRA) for six

human cell lines. Further, *GeneHancer* (Fishilevich *et al.*, 2017) combines four different REM databases and aims to reduce redundancy within the data. The *EnhancerAtlas* might be the database including the highest number of experimental methods. Applying an unsupervised learning approach, consensus enhancers are identified based on 12 different data types.
To understand the regulatory function of a REM one needs to know not only the REM location, but also the associated target gene. One possibility is to link REMs to their nearby gene. For instance, *Vista Enhancer Browser* offers the functionality to report REMs within a defined window up- and downstream of a candidate gene without taking into account any additional information. *RAEdb* provides REMs localized in the promoter regions of the gene of interest. However, REMs can be far away from their target genes and still interact with the promoter region by forming loop structures (Sanyal *et al.*, 2012; Jeong *et al.*, 2008). Therefore, all other resources mentioned above link the REMs to genes with different approaches considering additional genetic information. For instance, the *FANTOM5 Human Enhancers* atlas associates REMs to genes computing a pairwise expression correlation between the REMs eRNA and the promoters of candidate genes. *HACER* uses the associations from different chromosome conformation capture techniques and experimentally validated interactions to link the identified REMs to genes. The *GeneHancer* database determines a score for the REM-gene interactions by combining Hi-C data with eQTLs and co-expression correlation of enhancers and genes. EAGLE (Gao and Qian, 2019a) is a method that determines REM-gene interactions based on six genomic features, and is used to link the REMs in the *EnhancerAtlas* to genes.
Further, there are databases which are more focused on the impact of REMs in human diseases and their potential disease-related target genes (Wang *et al.*, 2017; Zhang *et al.*, 2017).

4.1.2 Using *STITCHIT* to generate data hosted by EpiRegio

EpiRegio utilizes regulatory elements predicted by *STITCHIT* (Schmidt *et al.*, 2021), an algorithm mainly developed by Florian Schmidt and Alexander Marx summarized in Section 2.2.6. In comparison to the methods mentioned above which usually first determine the REMs and in a second step associate them to target genes, *STITCHIT* performs the identification of the REMs and the linking simultaneously and thereby avoids peak calling.
We applied *STITCHIT* to human paired DNase1-seq and RNA-seq data on 110 samples taken from the Roadmap consortium (Kundaje *et al.*, 2015) and 56 samples from the Blueprint consortium (Stunnenberg *et al.*, 2016) (see also Section 2.1.8). The Roadmap data consists of a variety of cell types and tissues, while the Blueprint data is mainly derived from primary cell types and

disease associated samples of the haematopoietic system. We uniformly preprocessed the data, adjusted the DNase1-seq data for sequence depth and quantified the gene expression in TPM (see Section 2.2.2). For each annotated gene *STITCHIT* learned a separate model to identify the REMs based on a user-defined region around the gene. To identify even distal REMs, we predicted them within a window of 100,000 bp upstream of the gene's TSS, the entire gene body and 100,000 bp downstream of the gene's transcription termination site (TTS). *STITCHIT* learns a model per gene for each dataset. Thus, it can happen that for a gene REMs are predicted from both the training on the Roadmap data, as well as from the Blueprint data. In such cases we decided to take the predicted REMs only from the Roadmap consortium since the data is more diverse and resulted in a higher prediction performance.
As a result, 2,404,861 REMs associated to 35,379 protein-coding and non-protein coding genes, with an average length of 229 bp (+/- 235 bp) were identified. These REMs and associated target genes form the basis of the EpiRegio webserver. For each REM the regression coefficient and the p-value of the OLS model is provided, which we utilize within EpiRegio.

4.1.3 The EpiRegio webserver

Our EpiRegio webserver holds the REMs and their associated target genes, annotated using *STITCHIT* for 33 tissues and cell types from the Roadmap consortium and for 13 cell types from the Blueprint consortium. Three different queries are possible: (1) The *Gene Query* to determine the associated REMs for given genes, (2) the *Region Query* to extract REMs within a genomic region and (3) the *REM Query* to identify to which gene a specific REM is linked to. The result for each query is provided in an interactive table. An overview is visualized in Figure 4.1.

4.1.4 The three types of queries and their required input

Given one or several Ensembl IDs or gene names, EpiRegio's *Gene Query* returns a list of all REMs linked to the considered genes. The *Region Query* takes genomic regions in the format *chr:start-end* and returns all REMs which overlap. The *REM Query* addresses users who are already familiar with our database. For each given REM ID, the linked target genes are returned.
The input to each query can either be typed into our web interface or for large analyses uploaded as *.csv* or *.txt* file. In case of the *Region Query* a *.bed* file is additionally accepted. For each query it is possible to provide a list of celltypes or tissues of interest. When doing so, the user retrieves information about the REMs' contribution to the target genes' regulation and the activity of the identified REMs for the given cell types or tissues. The activity and the

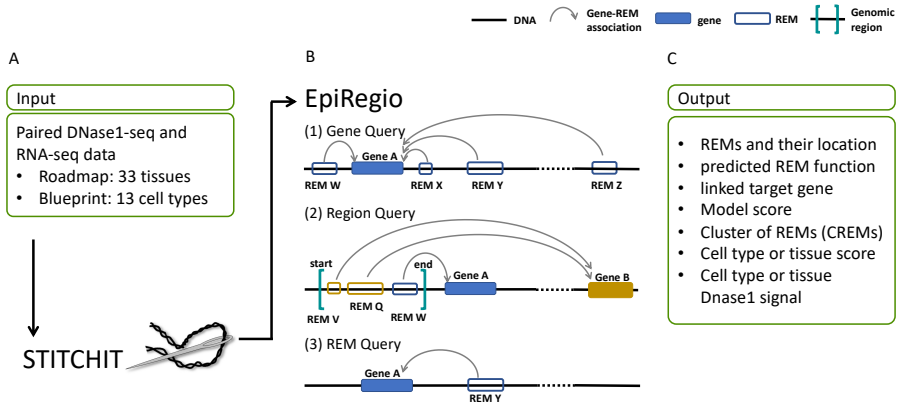


Figure 4.1: Overview of the structure of EpiRegio. (**A**) Human paired DNase1-seq and RNA-seq data from the Blueprint and Roadmap consortium were used as input to STITCHIT, which predicts REMs and their linked target genes. (**B**) Our EpiRegio database handles three types of queries. The user can either (1) identify REMs for a given gene, (2) check which REMs fall within a genomic region and (3) look up REMs based on their ID. (**C**) The query result is an interactive table containing various kind of information, which are explained in detail in Section 4.1.5.

contribution on the genes' regulation are represented as scores. One can specify thresholds on these scores such that only REMs exceeding them are returned.

4.1.5 Output of the interactive result table

The result of each EpiRegio query is an interactive table holding the information of the associated genes. In Figure 4.2 an example is illustrated. In the following the information given in the table for each REM are explained in detail.

Gene ID, REM ID and genomic location

Each row in the result table gives the *Gene ID* and the corresponding *Gene symbol* based on GRCh38.p10. Each REM is associated to a unique *REM ID*. Even if two REMs have the identical genomic location linked to different genes, the REM ID is a different one. Further, the genomic position of a REM is given in the columns *Chr*, *REM start* and *REM end*.

Results based on your query for the genes: DBR1

Gene ID	Gene symbol	REM ID	Chr	REM start	REM end	Predicted function	Model score	Cluster of REMs (CREM) ID	Number of REMs in the CREM	Heart score (n=2)	Heart DNase1 signal (n=2)
ENSG00000138231	DBR1	REM1643186	chr3	138061012	138061951	repressing	0.5038420	CREM0258490	2	-0.0459012	1.1149245
ENSG00000138231	DBR1	REM1643189	chr3	138062152	138062301	activating	0.5298390	CREM0258491	2	0.0290645	2.353615
ENSG00000138231	DBR1	REM1643191	chr3	138062802	138062901	activating	0.5553150	No CREM	-	0.0930812	7.35882
ENSG00000138231	DBR1	REM1643194	chr3	138063702	138064401	repressing	0.1541190	CREM0258492	2	0.0019201	0.1939475

Figure 4.2: Exemplary result table of a *Gene Query*. The first few rows of the result table of the *Gene Query* for the gene DBR1 are shown.

REMs' predicted function and impact on regulation

Based on the regression coefficient *STITCHIT* provides, we can conclude how the REM is affecting the target gene's expression. A positive regression coefficient is associated to an activating effect of the REM, whereas a REM with a negative coefficient is assumed to have a repressive function on the target gene's expression. In the column *Predicted function* the effect of a REM is denoted by either *activating* or *repressing*.
The *Model score* of a REM is derived from the p-value of the regression coefficient provided by the *STITCHIT* model, describing the significance of the contribution of a REM for the target gene's expression. We took the absolute binary logarithm of the p-values of all REMs associated to a given target gene. To allow a better interpretation of the *Model score*, we normalized it between $[0, 1]$ by dividing by the highest observed value for all REMs linked to the considered target gene. The *Model score* can be used to interpret how important a REM is in relation to all other REMs of the target gene. The higher the score, the higher the impact on the expression of the target gene. This score can be used to rank the scores, but not in a cell type or tissue specific manner.

Cluster of REMs (CREM)

The *STITCHIT* model is applied to each gene separately, resulting in a separate set of REMs for each gene. Due to the gene-specific segmentation, REMs associated to different genes can overlap each other. To allow the investigation of regions with high regulatory potential across multiple genes, we provide Cluster of REMs (CREMs). CREMs are defined as regions consisting of overlapping or adjacent REMs and have a unique ID. The column *Cluster of REM (CREM) ID* either holds this ID if the current REM is part of a CREM or the term *No CREM* if the REM is not overlapping or adjacent to another REM. Further, the number of REMs within a CREM is given in the column *Number of REMs in a CREM*. By clicking on the CREM ID, one is redirected

to a table containing all REMs and their associated genes of the current CREM.

Optional cell type or tissue specific output

For each selected cell type or tissue, two additional columns are included in the result table: the *Cell type or tissue score* and the *Cell type or tissue DNase1 signal*. where *Cell type or tissue* is replaced by the given cell type or tissue. So, if the tissue *heart* is selected, the columns are denoted as *Heart score* and *Heart DNase1 signal*.
To estimate the REMs' contribution to the target gene's regulation in a given cell type or tissue, we define a *Cell type or tissue score*. This normalized score in the range $[-1, 1]$ represents the relative contribution of a REM r to its target gene expression in a cell type or tissue c:

$$\text{Cell type score}(r, c) := \frac{\beta_r \cdot \text{DNase1-signal}_{r,c}}{\sum_{r_i \in R} |\beta_{r_i} \cdot \text{DNase1-signal}_{r_i,c}|}. \tag{4.1}$$

β_r is the regression coefficient derived by the *STITCHIT* model, describing the impact of a REM to its target gene's expression in a non cell type or tissue specific manner. The DNase1-signal of a REM for a cell type or tissue is log-transformed and standardized for each REM over all cell types and tissues (mean= 0, standard deviation= 1). We normalize the contribution of a REM to its target gene ($\beta_r \cdot \text{DNase1-signal}_{r,c}$), by dividing it by the absolute sum of the cell type or tissue specific contributions of all REMs associated to this target gene. Thus, R can be defined as $\{r_1, ..., r_n\}$, where n is the number of REMs linked to the target gene of the REM r. By design, the absolute *cell type or tissue scores* of the REMs associated to a target gene sum up to 1. When multiple samples of a cell type or tissue are available, we average the DNase1 signal.
With the *cell type or tissue scores* it is possible to analyze the impact of different REMs associated to multiple genes within a cell type or tissue. In addition, one can compare the contribution of REMs associated to a target gene between cell types and tissues. A positive *cell type or tissue score* is expected to increase the target gene expression compared to the other cell types or tissues within our database and can be caused by the following two combinations of the regression coefficient β_r of the studied REM and the DNase1 signal:

1. β_r and $\text{DNase1 signal}_{r,c}$ are > 0, then the REM can be interpreted as active enhancer. Thus, the REM is likely to enhance the expression of the target gene in comparison to cell types or tissues where the chromatin is more closed.

2. β_r and DNase1 signal$_{r,c}$ are < 0, then the REMs are expected to function in a repressive manner, however, the chromatin is rather closed. This means the RM can not exert it's repressive effect which potentially leads to a higher gene expression.

On the other side, a negative score is predicted to have a decreasing effect on the target gene's expression compared to other cell types or tissues. A negative score can be observed in the two following ways:

1. $\beta_r < 0$ and DNase1 signal$_{r,c} > 0$, then the target gene's expression is decreased in comparison to other cell types or tissues within our database. Thus, such a REM can be interpreted as an active repressor.
2. $\beta_r > 0$ and DNase1 signal$_{r,c} < 0$, then the REM can be interpreted as an inactive activator. As a consequence, the gene expression is decreased in comparison to other cell types or tissues within EpiRegio.

The *Cell type or tissue DNase1 signal* provided in our result table is denoted as average log_2(DNase1-signal) of all samples of a cell type or tissue considered from the Blueprint and Roadmap database. It describes how accessible the chromatin of a REM is and therefore how active the REM might be. The DNase1 signal is normalized for the sequence depth and can be compared between different cell types or tissues.

4.1.6 Systematic access to EpiRegio via a REST API

In addition to the possibility to query EpiRegio via our web interface we provide a REST API based on Django's REST framework to allow a more systematic access to our data and to incorporate EpiRegio in pipelines and workflows. It can either be accessed in the browser or via a program that allows to make HTTPS requests, like the easy-to-use Python package *request*.
Our REST API allows three queries denoted as *GeneQuery*, *RegionQuery* and *REMQuery* providing similar results as the corresponding queries of the web interface. Additionally, we provide a query to retrieve all REMs within a CREM (*CREMQuery*), and a query to get detailed information of a gene (*GeneInfo*), like the genomic position, the gene symbol or the strand. All these queries follow the same syntax rules:

```
https://epiregio.de/REST_API/<query>/<optionalParamter>/<input>/
```

, where *query* can be *GeneQuery*, *RegionQuery*, *REMQuery*, *CREMQuery* or *GeneInfo*. Some queries take additional parameters, and as *input* the genes, REMs, CREMs or regions of interest should be given, separated by a underscore. The result is provided in JSON format. To retrieve the REMs within the CREM *CREM0000002* the following HTTPS request is used:

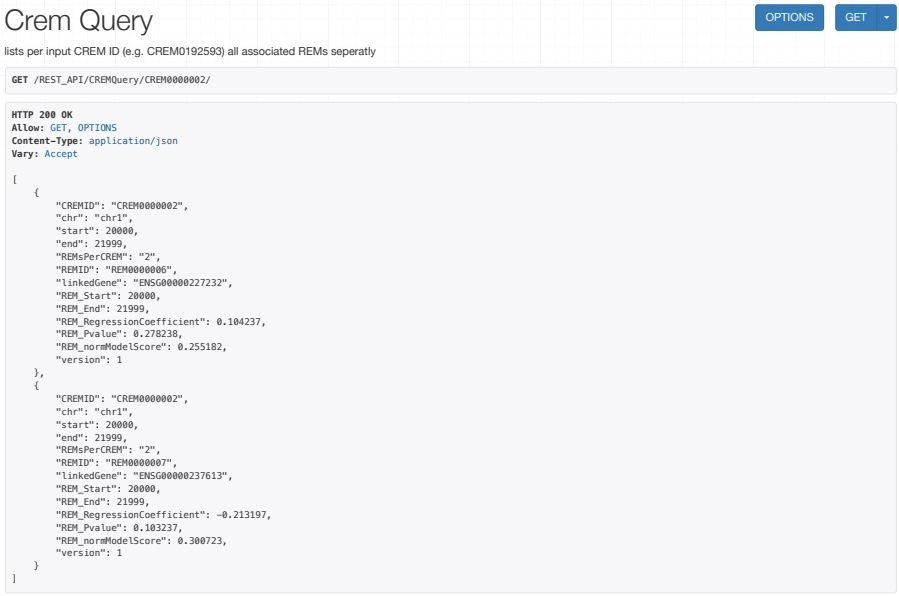


Figure 4.3: Example output of EpiRegio's REST API for a *CREMQuery*. The resulting output of the *CREMQuery* for the CREM *CREM0000002* based on Djangos REST framework is shown.

https://epiregio.de/REST_API/CREMQuery/CREM0000002/.

The resulting output is depicted in Figure 4.3. Examples for all possible query types as well as an example how to programmatically assess our REST API via Python are outlined in Appendix A.2.1.

4.1.7 System setup and structure of the database

The EpiRegio webserver was developed using the Python-based web framework Django (version 2.2.10, Python 3.7), where we created the result tables with the jQuery (version 1.19.1) library DataTables (version 1.10.20). The REST API is based on Djangos REST framework. Public access is provided by the Nginx (version 1.17.9) proxy service with Gunicorn (version 20.0.0) as the gateway interface. The source code is accessible at https://github.com/Team Regio/EpiRegioDB. The data hosted by our EpiRegio webserver is stored as a MySQL (v 8.0.19) database and publicly available at ZENODO (see https://doi.

org/10.5281/zenodo.3750929) (see also Appendix Section A.5.1). By providing our code as open-source and the data of EpiRegio as freely available, we ensure reproducibility.

4.2 Application: Computational prediction of CRISPR-impaired non-coding regulatory regions

In a cooperation with the lab of Manuel Kaulich, we analyzed CRISPR-induced mutations in the non-coding genome and developed an analysis protocol to identify target genes and to perform functional analyses based on the identified REMs. The work is published as part of the issue *Bioinformatics in theory and application* of *Biological Chemistry* with the title *Computational prediction of CRISPR-impaired non-coding regulatory regions* (Baumgarten *et al.*, 2021). Parts of the text and figures are taken from the publication or similar to it. The first authorship is shared between Florian Schmidt and myself. The contributions of all authors of the paper is as follows:

- Computational analyses for the identification of CRISPR impaired REMs and the functional analyses, figures: myself
- Developing analysis protocol, literature search and writing the text of the paper: Marcel H. Schulz (M.H.S) and myself
- General advice regarding CRISPR technology, text sections of CRISPR technique, their screen and parts of the introduction, providing of the non-coding gRNAs: Martin Wegner, Marie Hebel and Manuel Kaulich
- Data collection, *STITCHIT* analyses and background sampling: Florian Schmidt (F.S)
- Proofreading: all authors

When we started working on this paper, the EpiRegio webserver was not set up. Therefore, we took the REM-gene interactions directly from the *STITCHIT* publication. In comparison to EpiRegio's webserver, we considered all resulting models from the Roadmap and the Blueprint data, so it could be that for some genes the REMs of more than one model are included.

4.2.1 Overview of our analysis protocol and the used genome-wide CRISPR-Cas9 screen

Within this work, we propose an analysis protocol to associate non-coding target sites of a CRISPR screen to candidate target genes. Given the target regions of gRNAs, our analysis protocol identifies affected REMs, which are active in the studied cell type, and their associated candidate target genes. In a first step, we intersect the binding sites of the gRNAs with the REMs determined with the *STITCHIT* algorithm to identify the genomically modified REMs.

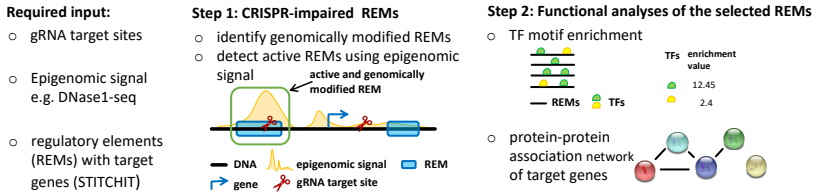


Figure 4.4: Overview of our analysis protocol. As input the gRNA target sites of a CRISPR screen, cell type-specific open chromatin data and REM-gene interactions are expected. Step 1: gRNA target sites are intersected with a catalog of REM-gene interactions and peak regions of epigenomic measurements of open chromatin (*e.g.* DNase1-seq, H3K4me3). Step 2: functional analyses of the associated REMs and target genes. A TF motif enrichment in REMs is performed and a protein-protein interaction network of the target genes is computed.

Further, cell type-specific epigenomic data for open chromatin, like DNase1-seq or histone modifications associated with active transcription *e.g.* H3K27ac or H3K4me3 is utilized. Only genomically modified REMs overlapping with open chromatin of the studied cell type are considered in further analyses. An overview is shown in Figure 4.4.
Within this work, we utilized a CRISPR-Cas9 screen in human telomerase-immortalized retinal pigmented epithelial (hTERT-RPE1) cells from Wegner *et al.* (Wegner *et al.*, 2019). The aim was to detect mutations that cause a chemotherapy resistance. The authors generated a truly genome wide (TGW) CRISPR library to target coding as well as non-coding regions of the human genome. The hTERT-RPE1 cells were induced with lentiviruses in such a way that per cell only one lentiviral integration event took place (multiplicity of infection of 1). Thereafter, the cells were treated with doxorubicin for 3 weeks. The surviving and presumably resistant cells were sequenced to detect the enriched gRNAs, which were then mapped onto the human reference genome to identify the modified regions. To ensure the quality of the gRNAs, the experiment was repeated with a new validation library only containing the enriched gRNAs. 795 gRNAs were re-identified and showed a high enrichment to doxorubicin treatment in comparison to controls not treated with doxorubicin. Several of the gRNAs affect protein coding regions, for which the authors could show that they are most likely linked to doxorubicin resistance.

4.2.2 Identification of REMs and TFs that mediate chemo-resistance in hTERT-RPE1 cells

To illustrate our analysis protocol on a realistic application, we utilized 332 non-coding target sites of 226 gRNAs identified within the CRISPR screen

from Wegner *et al.* (2019) in hTERT-RPE1 cells. In order to identify genomically modified REMs and their target genes, we first extended the target sites to 200 bp long regions such that the target site is centered in the middle. Next, we intersected these regions with $3,900,708$ REMs predicted by *STITCHIT* based on data from Blueprint, Roadmap and ENCODE (The ENCODE Project Consortium, 2012). The epigenetic data used to predict the REM-gene interactions was neither from hTERT-RPE1 cells nor from RPE-1 cells.
219 REMs linked to 190 different genes were identified to overlap with a non-coding gRNA. In order to verify that we did not observe the overlap by chance, we shuffled the positions of the target sites randomly within the genome and counted how many of these regions intersect with our REMs. After repeating this 100 times, we observed on average 72.07 REMs overlapping for the Blueprint consortium, 83.87 for Roadmap and 25.45 for ENCODE. We observed a significant enrichment for the overlap of the true gRNAs for Blueprint ($89/72.07$, p-value $< 2.2e-16$), Roadmap ($90/83.87$, p-value $1.638e-06$) and ENCODE ($40/25.45$, p-value $< 2.2e-16$) compared to the randomly sampled regions (two sided t-test).
While *STITCHIT* infers the REMs based on a broad variety of cell types and tissues, the exact cell type we are working on is not included. Thus, to ensure that we only consider REMs active in hTERT-RPE1 cells, we utilized open chromatin regions from DNase1-seq and H3K4me3 ChIP-seq data for hum PRE-1 cells downloaded from ENCODE. In Table 4.1 the 13 identified gRNAs target sites overlapping active REMs associated to 35 different target genes are shown. We noticed that several gRNA target sites intersect with multiple REMs linked to multiple genes. Thus, we conclude that these genes, which might be affected by changes in gene regulation induced by the genetically modified REMs, may be strong candidates to play a role in doxorubicin chemoresistance.
To provide evidence that our identified candidate genes are involved in chemoresistance, we conducted a detailed literature survey. We aimed to find previously published studies that link our candidate genes to chemoresistance against doxorubicin or other cancer drugs as well as to cancer growth, that might be involved in establishing chemoresitance. We were able to provide literature evidence for at least one candidate gene for 8 out of 13 gRNAs. In the Appendix A.2.3 the findings are explained in detail grouped according to the genes predicted to be affected by the same gRNA.
Since many genes from Table 4.1 are associated with chemoresistance, we analyzed if they form function modules within the cell. Thus, we used the STRING database (Szklarczyk *et al.*, 2020) to compute a customized protein-protein association network for the 35 candidate target genes. Additionally, we used the functionality of the STRING database to automatically include 10 proteins which are highly connected to the candidate genes (see Appendix Figure 1).

We found that the genes DOC2A, TMEM219, GDPD3 and TBX6 form an association module. These four genes lie within a region of 600 kb, which is connected to different forms of human diseases depending on whether the region was deleted or duplicated (Zheng *et al.*, 2014; Arbogast *et al.*, 2016).

chr	gRNA coordinate	gene(s) associated
1	40,036,652	CAP1, PPT1, AL663070.1
10	103,282,011	**NT5C2** (Dieck and Ferrando, 2019)
10	3,197,670	**PFKP** (Shen *et al.*, 2020)
10	73,358,465	CFAP70, RPL26P6, **ANXA7** (Liu *et al.*, 2020; Gao *et al.*, 2020), RNU6-833P
13	87,671,102	SLITRK5
16	30,021,011	**TMEM219** (Cai *et al.*, 2020), **DOC2A** (Indermaur *et al.*, 2010), TBX6, **GDPD3** (Baras *et al.*, 2015; Indermaur *et al.*, 2010)
17	28,404,283	AC002094.1, KRT18P55, SEBOX, **TMEM199** (Buckley *et al.*, 2019)
17	42,423,866	MIR5010
17	81,635,498	MRPL12, **SLC25A10** (Rochette *et al.*, 2020; Wang *et al.*, 2020), AC137896.1, HGS, TSPAN10, AC139530.3, CCDC137
20	45,993,823	**NCOA5** (Sun *et al.*, 2017), **MMP9** (Spallarossa *et al.*, 2006), PCIF1
22	24,063,890	AC253536.3
4	139,651,636	AC112236.2, H3P16, **MGST2** (Dvash *et al.*, 2015)
8	120,445,481	MTBP, COL14A1

Table 4.1: Candidate doxorubicin chemoresistance genes. All gRNAs that overlap active regulatory regions and their associated target genes are listed. In bold the genes are marked which have been reported in the literature to be linked to chemoresistance or cancer growth.

To further gain confidence in our findings, we performed a functional interpretation and annotation of the 35 candidate genes utilizing the David workflow (Jiao *et al.*, 2012). Using their functional annotation clustering functionality, 10 genes (SLITRK5, CAP1, DOC2A, GDPD3, HGS, MGST2, SLC25A10, TSPAN10, TMEM199 and TMEM219) were identified to be associated with membrane. Membrane transporters are known to be involved in the development of chemoresistance. An example is the ABC membrane transporter which can pump diverse cancer drugs out of the cells leading to a chemoresistance (Choi, 2005).
In order to identify TFs which might be involved in the regulatory process of doxorubicin resistance, we applied a TF motif enrichment analysis based on the active, genomically modified REMs. The analysis was performed using PASTAA (see Section 2.2.4). As input we provided the 45 active REMs affected by the gRNA target sites and 515 human TF binding motifs downloaded from the JASPAR database. In addition, a ranking of the REMs is required. Thus, we sorted the REMs in descending order based on their maximal epigenetic signal either from DNase1-seq data or the H3K4me3 ChIP-seq. Since 45 sequences are rather little to discriminate enriched TF binding sites and to get a reliable result with PASTAA, we added 3 times more randomly selected active REMs in RPE1 cells not affected by a gRNA. We sorted the background REMs in ascending order based on their epigenetic signal, and added them to the bottom of the ranking of the active and genomically modified REMs. The motif enrichment analysis was repeated 100 times, where for each run a different background set was utilized. The 20 most enriched TFs are shown in Table 4.2, out of which some were already shown to be associated to chemoresistance (marked in bold). In the Appendix A.2.5 we outline the TFs for which we found literature evidence.
We conclude, that our analysis protocol identified several interesting candidate genes and TFs which might be worth investigating in order to better understand the doxorubicin resistance in hTERT-RPE1 cells.

4.2.3 Analyses of further CRISPR-Cas screens also highlight target genes with literature evidence

In addition, we analyzed two CRISPR-Cas screens from Klann *et al.* (2017). These screens are focused screens that identify gRNA target sites associated to a significant gene expression change of the genes of interest. We chose them because of the lack of further suitable genome wide screens.
Since the analysis is similar as for the previous CRISPR-Cas screen, the details can be found in Appendix A.2.6. We were able to provide literature evidence supporting the usefulness of our analysis protocol for these focused screens as well.

rank	TF	FDR
1	**AP-1** (Wang *et al.*, 2018b)	3.83e-05
2	**NR2C2** (Hu *et al.*, 2019)	8.14e-05
3	CREB3L4	0.00015
4	HAND2	0.00016
5	**ATF3** (Nobori *et al.*, 2002; Park *et al.*, 2012)	0.00017
6	**ATF7** (Walczynski *et al.*, 2013)	0.00018
7	MYBL1	0.00035
8	E2F8	0.00038
9	**ATF2** (Walczynski *et al.*, 2013)	0.00039
10	NR2F1	0.00059
11	RARA	0.00062
12	NR3C1	0.00062
13	DBP	0.00079
14	**NR2F2** (Wang *et al.*, 2016)	0.00084
15	NR2C1	0.00090
16	GMEB2	0.001
17	THRB	0.001
18	**SREBF2** (Wang *et al.*, 2019)	0.0012
19	**RUNX3** (Zheng *et al.*, 2013)	0.0013
20	**SREBF1** (Wang *et al.*, 2019)	0.0013

Table 4.2: Result of the TF enrichment analysis. TF binding in genomically modified active REMs was analyzed in comparison to random sets of active REMs in the same cells. The ranked TF names and their FDR corrected enrichment p-values are shown. In bold marked are the TFs which have been reported to be linked to chemoresistance.

4.2.4 Conclusion

To conclude, we introduced a cell type-specific analysis protocol that allows to identify potential target genes of CRISPR-induced modifications in the non-coding genome. We utilized the predicted REM-gene interactions of *STITCHIT*.

The analysis is also possible with EpiRegio's *Region Query*. Based on EpiRegio's *cell type and tissue signal* one can directly retrieve active REMs from the webserver. In addition, we outlined easy to apply downstream analyses based on the identified REMs like the TF motif enrichment analysis.
For the genome-wide CRISPR screen we provide literature evidence that the identified target genes and TFs are known to be linked to doxorubicin resistance, chemoresistance or cell growth in cancer cells. Our analysis protocol based on epigenetic data might be helpful to point to novel target genes and identify TF that regulate those genes for different CRISPR screens and various cell types.

4.3 Application: Circular RNA circPLOD2 regulates pericyte function by targeting the transcription factor KLF4

The following work is a cooperation with the labs of Kathi Zarnack and Stephanie Dimmeler. Our contribution was to point to the TF KLF4 which leads to circPLOD2-mediated expression changes. For that, we used EpiRegio's *Gene Query* followed by a TF motif enrichment analysis. On EpiRegio's GitHub repository we provide a computational workflow for this analysis. A similar analysis is also outlined in our EpiRegio documentation.
The paper *Circular RNA circPLOD2 regulates pericyte function by targeting the transcription factor KLF4* is available as preprint on bioRxiv (Glaser *et al.*, 2022). Parts of the following sections include the figures and text to which we contribute in this paper. In the following, the contribution of all authors with the focus on the bioinformatics analyses is listed:

- Identification of REMs associated to DEGs, TF motif enrichment analysis, and deeptools analysis: myself
- discussion and design of our bioinformatics analyses: Marcel H. Schulz (M.H.S) and myself
- Design of experiments: Christopher Zehendner (C.Z), Simone Franziska Glaser (S.F.G) and Stefanie Dimmeler (S.D)
- All experiments were done by S.F.G and Marius Klangwart (M.K)
- RNA-sequencing was performed by Stefan Günther
- Quantification of circRNAs, RNA-seq and DEGs analysis on circPLOD2 depleted and control pericytes: David John (D.J), Andre Brezski (A.B) and Kathi Zarnack (K.Z)
- miRNA analysis: Ranjan Kumar Maji (R.K.M)
- Writing the paper: S.F.G, S.D. A.B, K.Z, R.K.M, M.H.S and myself,
- Proofreading: all authors

The following part is divided in two sections: The first section summarizes the findings that motivated us to analyze which TFs might be involved in the circPLOD2-mediated expression changes. The second section outlines our analysis in detail.

4.3.1 Identification of circPLOD2 under hypoxia conditions in human pericytes

In this study we aimed to characterize the expression and also the regulation of circular RNAs (circRNAs) in human vascular pericytes under hypoxia. Hypoxia mimics the dropping oxygen supply taking place during myocardial infarction. CircRNAs are a class of RNAs, usually translated from protein-coding genes, that form a closed circle via a backsplicing process (Memczak *et al.*, 2013; Jeck *et al.*, 2012; Salzman *et al.*, 2012). Both transcripts, the linear protein-coding one and the circular one can be detected within a cell. CircRNAs often behave like non-coding RNAs, and can fulfill various of functions in different tissues or cell types (reviewed in Qu *et al.* (2015b)).
Based on previously published RNA-seq data of human brain vascular pericytes exposed to reduced oxygen supply (Bischoff *et al.*, 2017), circPLOD2 was identified as most strongly upregulated among other circRNAs. Thus, we decided to study circPLOD2 further. Within the lab, we were able to confirm that the depletion of circPLOD2 affects pericyte functions in various ways, *e.g.* cell motility, endothelial-pericyte crosstalk via paracrine signaling and the stimulation of endothelial tube formation. To better understand the gene expression changes induced by circPLOD2 depletion, we performed RNA-seq of circPLOD2 depleted and control pericytes. A differential genes expression analysis resulted in $2,510$ differentially expressed genes. Within a functional enrichment analysis of the DEGs various cell cycle processes were enriched, confirming the previous findings observed in circPLOD2 depleted pericytes. In order to explain the gene expression changes, we investigated the role of miRNAs. However, we did not find an increased density of miRNA target sites nor a miRNA enrichment. Therefore, we conclude that circPLOD2 is most likely not regulated by miRNAs.

4.3.2 Changes in gene expression are mediated via regulation of the TF KLF4

To identify TFs involved in the gene expression changes observed for circPLOD2 depleted pericytes, we computed a motif enrichment analysis on REMs linked to DEGs. To point to a specific TF, we visualized the number of binding sites of the enriched TFs in a heatmap.
In a first step, we utilized EpiRegio's *Gene Query* to collect REMs associated

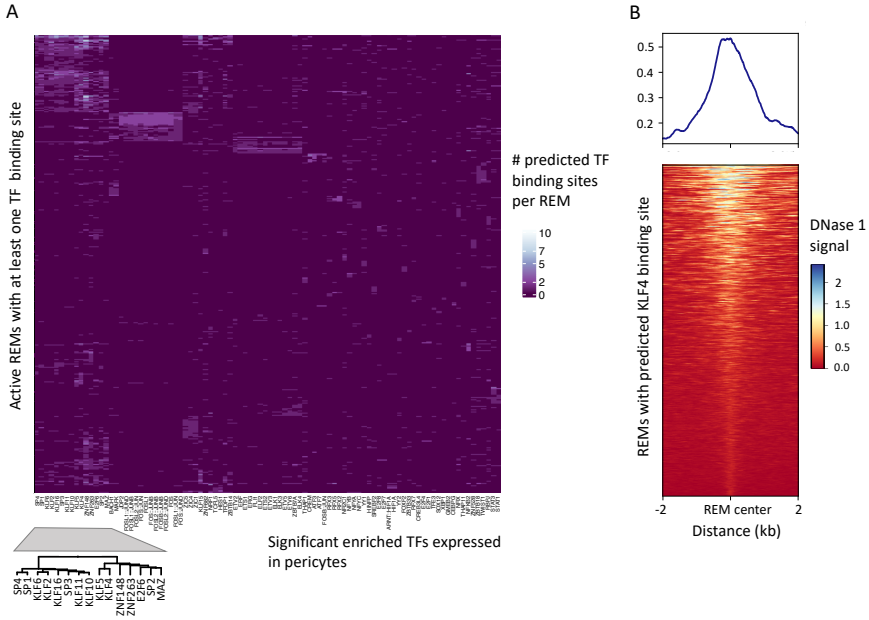


Figure 4.5: Identification of candidate TFs involved in regulating DEGs. (**A**) The heatmap visualizes the number of predicted TF binding sites per REM (y-axis) for 95 significant enriched TFs (x-axis) with the color indicating the number of binding sites (the lighter the more predicted binding sites). (**B**) An average profile (top) and a heatmap show the chromatin accessibility of REMs with a KLF4 binding site within a 4 kb window centered at the middle of the REMs (bottom).

to the top 1000 DEGs sorted according to their adjusted p-value. Since the studied cell type was not included in EpiRegio, we could not directly use the *cell type or tissue DNase1 signal* to filter for active REMs. Instead, we downloaded DNase1 signal of human brain pericyte from ENCODE and computed their activity with *STITCHIT*'s REMSelect functionality. We excluded those REMs not active in the studied cell type (mean DNase1 signal < 0.1), resulting in $16,901$ REMs. Using *PASTAA* (see Section 2.2.4), we computed a motif enrichment using 292 motifs of expressed TFs (TPM > 1) in pericytes. As input sequences we used the active REMs sorted in descending order according to their DNase1 signal. The analysis resulted in 95 significantly enriched TFs (adjusted p-value ≤ 0.01) (see Appendix A.2.7).

To further decrease the number of TFs and to point to candidate TFs that might have the largest effects on the differentially expressed genes, we counted the number of predicted binding sites for each significantly enriched TF over all REMs and visualized the counts in a heatmap (see Figure 4.5A). The TF

binding sites were predicted using Fimo (see Section 2.2.3). A particularly interesting cluster was formed by several members of the *Krüppel like factors* (KLF) which had many TF binding sites in a large number of REMs (Figure 4.5A upper left corner). We identified KLF4 as an interesting candidate, since it had at least one binding site in 3,025 active REMs associated to 767 out of 1,000 DEGs. Additionally, KLF4 was significantly downregulated in pericytes with circPLOD2 depletion. To confirm that investigating KLF4 might be reasonable, we analyzed the chromatin status of the REMs with at least one KLF4 binding site (see Appendix A.2.8). In Figure 4.5B one can observe the DNase1 signals of the REMs in control conditions. The open chromatin status of the REMs supports the theory that KLF4 may play a role in the regulation of the DEGs.
With an overexpression of KLF4 in circPLOD2 depleted pericytes, we were able to show that the effects of a circPLOD2 depletion are prevented.

4.3.3 Conclusion

To sum up, we illustrated an application of EpiRegio's *Gene Query* with a commonly used downstream TF inference analysis, which is usually the first step to unravel the regulatory pathways of the DEGs in the studied conditions. Thanks to the collaboration with the lab of Stefanie Dimmeler we were able to showcase that the identified TF KLF4 is involved in regulating the gene expression changes observed in circPLOD2 depleted pericytes.

4.4 Discussion

EpiRegio is a webserver holding about 2.4 million REM-gene interactions predicted by the unique approach of *STITCHIT*. The gene centric model analyzes variations in epigenetic data in relation to gene expression across different cell types and tissues and thereby identifies REMs and their associated target genes simultaneously. In comparison, other approaches often define the REMs first based on CAGE or epigenomic data and then apply varying techniques to associate the REMs to genes (Wang *et al.*, 2018a; Andersson *et al.*, 2014; Fishilevich *et al.*, 2017).
Further, we want to mention that *STITCHIT* uses a linear model to retrieve the regression coefficient. Clearly this is a simplification of how REMs regulate their gene's expression. The linear model assumes that the contribution of each REM is not only independent, but also additive, which might not be true for all of them and is under debate (Dukler *et al.*, 2017). Nevertheless, we hope that the score is useful for many applications.
With our EpiRegio webserver we provide an easy-to-access tool to the com-

munity to efficiently retrieve REMs and their linked target genes. We aim to overcome the use of simplistic methods like linking a REM to its nearest gene. The different ways to access our webserver are intuitive to use and allow for various kinds of queries. Further, we provide a REST API which enables users to access data programmatically.
We show the usefulness of EpiRegio in two applications. We first analyzed a CRISPR screen and associated non-coding gRNA target sites to candidate genes. This can be done by using for instance EpiRegio's *Region Query*. Additionally, we took epigenetic data into account to identify those genomically modified REMs which are located in open chromatin or active enhancer regions in RPE-1 cells. By utilizing an intensive literature search, we provided evidence that the candidate genes associated to active and genomically modified REMs are linked to drug resistance in cancer. As additional downstream analysis we performed a TF motif enrichment analysis, where we pinpoint several TFs involved in drug resistance in cancer or cancer growth according to literature. Further, we constructed a protein-protein association network of the candidate genes and performed a functional interpretation and annotation of these genes.
In our second analysis, we helped to identify which TFs might be involved in regulating differentially expressed genes in circPLOD2 depleted pericytes. For that, we used EpiRegio's *Gene Query* to collect REMs linked to DEGs. A downstream TF motif enrichment analysis resulted in a list of candidate TFs Additionally, we provided a heatmap visualizing the counts of predicted TF binding sites per REM for each enriched TF, which allowed us to identify the TF KLF4 as particularly interesting. Within the lab it was shown that KLF4 is involved in the expression changes observed for circPLOD2 depleted pericytes.
To sum up, we believe that EpiRegio is a useful and accessible webserver, as we demonstrated in the two applications outlined in detail above. Besides the presented examples, EpiRegio can be used for a variety of scientific questions and serves as a source of information that is relevant in many different scenarios, like understanding gene regulatory networks, identifying target regions for experimental setups or exploring candidate genes of TFs.

Chapter 5

A statistical approach to identify regulatory DNA variations

In this chapter, our statistical approach to identify regulatory Single Nucleotide Polymorphisms (SNPs) that affect the binding sites of Transcription Factors (TFs) is introduced. It is based on our preprint *A statistical approach to identify regulatory DNA variations* which is publicly available on bioRxiv (Baumgarten *et al.*, 2023). Therefore, most of the text and the figures are taken from the preprint. In the following, the contributions of all authors are listed

- Necessary computational steps, code development, figures and writing the paper: myself
- General discussion regarding the approach, the results and the paper, developing the ideas of the approach, finding suitable datasets to evaluate the approach, literature search: Marcel H. Schulz and myself
- Set up bioconda SNEEP package: Laura Rumpf
- Evaluation of the SNEEP result for atherosclerosis GWAS, writing the corresponding section in the text: Thorsten Kessler (data not shown within this Chapter)
- Proofreading: all authors

5.1 State of the art methods in comparison to our new approach

Non-coding Single Nucleotide Polymorphisms (SNPs) localized in regulatory elements, such as enhancers or promoters can lead to changes in gene expres-

sion by modifying Transcription Factor Binding Sites (TFBS) (see Background Section 2.1.7). Several studies reported the resulting functional consequences (reviewed for instance in Zhang and Lupski (2015)). Therefore, methods pinpointing to such regulatory SNPs (rSNPs) are a topic of current interest.
Several existing methods successfully highlight rSNPs based on epigenetic information, such as open chromatin data, TF and histone ChIP-seq without taking into account which TF might be affected (Lee *et al.*, 2015; Kelley *et al.*, 2016; Amariuta *et al.*, 2019; Chen *et al.*, 2022).
In contrast, methods that evaluate the effect of a SNP on a TFBS rely on the ability to describe the binding behavior of a Transcription Factor (TF) to assess the difference induced by a non-coding SNP. TFs are able to bind the DNA by recognizing short patterns. These patterns can be described *in vitro* using high throughput methods like protein-binding microarrays (PBM) (Berger *et al.*, 2006) or SELEX (Jolma *et al.*, 2010), or *in vivo* using ChIP-based techniques (Lambert *et al.*, 2018; He *et al.*, 2015). The identified TF binding preferences are summarized in a TF model, of which the most prominent are Position Weight Matrices (PWMs) (Stormo, 2000). More details about other existing TF binding models can be found in Section 2.2.3.
Computational approaches have been developed to evaluate the effect of a SNP on the binding sites of a TF. Among them is for instance *GERV* (Zeng *et al.*, 2015), a k-mer based approach that learns *de novo* TF binding based on open chromatin and TF ChIP-seq data. To evaluate the impact of a SNP they compute the difference of the predicted read counts for the two allelic variants of a SNP. A more recently published method is *FABIAN-variants* (Steinhaus *et al.*, 2022). They determine a differential TF binding score not only based on PWMs, but also on TFFMs and allow to take TFBS from epigenetic data into account. In comparison, QBIC-Pred (Martin *et al.*, 2019; Zhao *et al.*, 2017) utilizes *in vitro* universal PBM data to determine the TF binding behavior with a k-mer based model using ordinary least squares (OLS). They score the effect of a SNP based on the parameters of the OSL z-score, and additionally provide a p-value for their score. Further, there are methods like *sTRAP* (Manke *et al.*, 2010), *is-rSNP* (Macintyre *et al.*, 2010) or *atSNP* (Zuo *et al.*, 2015) that rely solely on TF models (usually PWMs) and the DNA sequence itself. To evaluate the effect of a SNP on TF binding sites different statistical approaches are introduced by these methods. *sTRAP* ranks TFs based on the difference of the TF binding scores for the wildtype and the alternative allele of a SNP, whereas *is-rSNP* and *atSNP* provide a p-value for their differential TF binding score. These methods are explained in more detail in Section 2.2.7.
In general, we noticed that only a few methods provide a statistical significance for their introduced differential TF binding scores. However, this is necessary to decide if a score is significantly different from the commonly assumed null hypothesis that the SNP does not affect the TF binding site. Further, if the TF

binding score is represented as p-value they are directly comparable between the TFs, which is often not possible for the scores itself. When the methods provide a p-value, their statistic is dependent on their underlying TF model. For instance, QBIC-Pred derives their test statistic based on OLS estimation of k-mers, whereas *atSNP* assumes the scores follow a multinomial distribution to model PWMs. *is-rSNP* is with the best of our knowledge the only method which allows to compute a p-value independent of the TF model. However, they determine the exact p-value distribution for differential TF binding scores by assessing all possible single base changes resulting in an algorithm scaling quadratically.
In this chapter, we introduce a fast and accurate approach to provide statistical significance for the differential TF binding score for general TF models. To do so, we examine the distribution of the maximal differential TF binding scores and find that it can be well approximated by a modified Laplace distribution. By using the modified Laplace distribution, we can derive a p-value for the maximal differential TF binding score in constant time. We show on experimentally validated TF-SNP pairs that our approach improves the performance in comparison to the previously established method *atSNP*, while being an order of magnitude faster.

5.2 A distribution fitting the maximal differential TF binding scores

In the following, we explain how to evaluate the effect of a non-coding SNP on a TFBS. In the first section, we outline how to compute the differential TF binding score for a fixed TF position, termed *The differential TF binding problem*. Next, we describe a more general case where the differential TF binding is computed for all sequences overlapping the SNP denoted as *The maximal differential TF binding problem*. Further, we explore and approximate the distributions of the differential TF binding score as well as the maximal differential TF binding score.

5.2.1 The differential TF binding problem

Let M be a general TF model of length m, $S = \{s_1, ..., s_m\}$ a DNA sequence of length m with $s_i \in \Sigma$ over the alphabet $\Sigma = \{A, C, G, T\}$ and assume there is a SNP with two allelic variants called wildtype and alternative allele at position i with $i \in \{1, ..., m\}$ in S. Hence, we consider two variants of the sequence S, S^1 containing the wildtype allele and S^2 the alternative allele. For each of these sequence variants we compute a TF binding score describing the TF binding affinity to the sequence according to the model M. From the distribution of

all binding affinity values for a TF under the model M, we can compute the probability $P(a \leq t, M)$, the p-value of observing a binding affinity a smaller than a threshold t. We can use this to determine the corresponding p-value of the model M for each variant, given as $p(S^1, M)$ and $p(S^2, M)$, respectively. The exact computation of the TF binding score and the p-value depend on the used TF model (see Section 2.2.3). To evaluate the effect of a SNP on a TFBS, we compute a log-ratio similar to Manke *et al.* (Manke *et al.*, 2010) in their method *sTRAP*, called *differential TF binding score (D)*:

$$D(S^1, S^2, M) = \log\left(\frac{p(S^1, M)}{p(S^2, M)}\right). \tag{5.1}$$

A positive D indicates that the exchange from the wildtype allele to the alternative allele increases the binding affinity of the TF and may lead to a gain of a binding site. Whereas a negative D decreases the binding affinity and therefore the binding site might be lost.

5.2.2 The maximal differential TF binding problem

Usually it is not known at which position the binding affinity is most affected by the SNP. As a consequence, we evaluate all sequences overlapping with the SNP, hence we define a window of size $2m-1$ centered around the SNP. We slide the TF model over the $2m-1$ long sequence and compute D for each of the sub-sequences of length m overlapping the SNP. To identify the optimal D, we keep the maximal absolute value.
Thus, our definition changes: We are given a DNA sequence $S = \{s_1, ..., s_{2m-1}\}$ and a SNP with two allelic variants centered in the middle of the sequence at position m. We consider two variants of the sequence S: S^1 containing the wildtype allele and S^2 including the alternative allele. Further, a k-mer which is a sub-sequence of S of length m, is defined as $k_i = \{s_i, ..., s_{i+m-1}\}$ with $i \in \{1, ..., m\}$. For each sequence variant, we define the k-mers overlapping the SNPs. The k-mers of S^1 are denoted as k_i^1 and the k-mers for S^2 as k_i^2 with $i \in \{1, ..., m\}$. We maximize over the absolute D values of all k-mers overlapping the SNP and identify the TF position where the binding site is most affected. The *maximal absolute TF binding score* D_{max} is defined as:

$$D_{max}(S^1, S^2, M) = \max_{i \in \{1,...,m\}} (|D(k_i^1, k_i^2, M)|). \tag{5.2}$$

5.2.3 Estimating the distribution of differential TF binding scores

In this section, we investigate if the distribution of D follows a known distribution. To do so, we describe the p-values of the TF binding scores for the sequences S^1 and S^2 of length m as random variables W and Z, respectively.

Since the TF binding score itself is continuous, the p-values are uniformly distributed in [0,1] under the null hypothesis (Wasserman, 2010). Further, we assume that the random variables W and Z are independent of each other. Hence, the following theorem can be applied:

Theorem 5.1 *If two independent random variables W, Z, are uniformly distributed in the range $[0, 1]$ then $log(\frac{W}{Z})$ is $Laplace(0, 1)$ distributed (Kotz* et al.*, 2001, p. 24).*

In theory, we conclude that D can be represented as random variable $X = \log(\frac{W}{Z})$ which is $Laplace(0, 1)$ $(L(0, 1))$ distributed (see Section 2.2.1). However, our experiments will show that D is not following the $L(0, 1)$ distribution (see Section 5.4.1). Nevertheless, D can be approximated by a $L(0, b)$ distribution, with a scale parameter b different from 1. The scale parameter is given as $b = \sqrt{\frac{\sigma^2}{2}}$, where σ^2 is the variance of the observed differential TF binding scores of a TF model M for a set of SNPs.

5.2.4 Derivation of the distribution of the maximal differential TF binding scores

When computing D_{max}, we can represent the differential TF binding scores of the k-mers as n independent and identical $L(0, b)$ distributed random variables X_i with $i \in \{1, ..., n\}$, where n is the overall number of k-mers. D_{max} can be described as the absolute maximum over all random variables X_i, so $Y = \max_{i \in \{1,...,n\}}(|X_i|)$. To efficiently compute a p-value for D_{max}, we are interested in identifying the cumulative distribution function (CDF) of Y.
To do so, we split the following in three parts: First we determine the probability distribution function (PDF) and CDF of the absolute values of a Laplace(0,b) distributed variable. Second, we derive the CDF of the maximum of n $L(0, b)$ distributed random variables, and finally, we combine both parts to get the CDF of Y.

Computation of the PDF and CDF of the absolute values of $L(0, b)$ distributed random variables

The general PDF of the $L(0, b)$ distribution with the scale $b > 0$ is defined as

$$f(x) = \frac{1}{2b} \cdot e^{\frac{-|x|}{b}}. \tag{5.3}$$

To obtain the density for the absolute value $|x|$, one needs to add up the densities for the positive and negative values:

$$f_{|x|}(x) = f(x) + f(-x) = \frac{1}{2b} \cdot e^{\frac{-|x|}{b}} + \frac{1}{2b} \cdot e^{\frac{-|-x|}{b}} = \frac{1}{b} \cdot e^{\frac{-|x|}{b}}, \tag{5.4}$$

with $x \in \mathbb{R}^+$. The corresponding CDF is given by integrating $f_{|x|}(x)$ from 0 to x:

$$F_{|x|}(x) = \int_0^x f_{|x|}(x)dy = \int_0^x \frac{1}{b} \cdot e^{\frac{-|y|}{b}} dy = \left[-e^{\frac{-|y|}{b}}\right]_0^x = 1 - e^{\frac{-|x|}{b}} \tag{5.5}$$

Derivation of the CDF of the maximal of n $L(0, b)$ distributed random variables

The CDF of a random variable V is defined as $F(x) = P(V \leq x)$. We derive the CDF of the maximal X_i with $i \in \{1, ..., n\}$ using the definition of CDFs and the independence of the $L(0, b)$ distributed random variables:

$$\begin{aligned} F_{max}(x) &= P(\max(X_i) \leq x) \\ &= P\left[(X_1 \leq x) \cap (X_2 \leq x) \cap ... \cap P(X_n \leq x)\right] \\ &= \prod_{i=1}^{n} P(X_i \leq x) = \prod_{i=1}^{n} F(x) = F(x)^n \end{aligned} \tag{5.6}$$

Derivation of the CDF for an maximal absolute $L(0, b)$ distributed random variable

To finally get the CDF of $Y = \max_{i \in \{1,...,m\}}(|X_i|)$, we can plug in the CDF for the absolute $L(0, b)$ distributed random variables (5.5) in Equation (5.6) resulting in:

$$F_{max|x|}(x) = F_{|x|}(x)^n = \left(1 - e^{\frac{-|x|}{b}}\right)^n. \tag{5.7}$$

The corresponding PDF is the derivative of Equation (5.7):

$$\begin{aligned} f_{max|x|}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}(F_{max|x|}(x))^n = \frac{\mathrm{d}}{\mathrm{d}x}(1 - e^{\frac{-|x|}{b}})^n \\ &= \frac{n}{b} \cdot e^{\frac{-|x|}{b}} \cdot (1 - e^{\frac{-|x|}{b}})^{n-1}. \end{aligned} \tag{5.8}$$

To sum up, we are able to mathematically describe the distribution of D_{max} as follows:

Theorem 5.2 *Let D_{max} be defined as $Y = \max_{i\in\{1,...,n\}}(|X_i|)$, where the X_i's are independent and identical $L(0, b)$ distributed random variables, then Y follows a modified Laplace distribution $L_{max}(n, b) = f_{max|x|}(x)$.*

The distribution of D_{max} is depending on the parameter n and the scale parameter b. Here, n denotes the number of k-mers overlapping the SNP times two, since we also consider the reverse complement.
To determine the scale parameter b for j observed maximal differential TF binding scores x_i with $i \in \{1, ..., j\}$ of a TF model M, we set up a maximum log-likelihood estimator (MLE) (see Section 2.2.1) of the PDF of L_{max}

$$\begin{aligned} L(x_i|b) &= \log\left(\prod_{i=1}^{j} f_{max|x|}(x_i)\right) \\ &= \log(k) - \log(b) - \sum_{i=1}^{j}\left(\frac{|x_i|}{b} + \log\left(1 - e^{\frac{-|x_i|}{b}}\right)^{n-1}\right) \end{aligned} \tag{5.9}$$

and computed the corresponding derivative with respect to b. Since the resulting equation is not analytically solvable we used Newton's method to numerically approximate b.
Using the CDF of the $L_{max}(n, b)$ distributed maximal differential TF binding scores, we are able to compute a p-value for D_{max} as $1 - F_{max|x|}(x)$.

5.2.5 Fitting the scale parameter b

In all analyses we conducted within this work we used as TF motif set a collection of non-redundant human PWMs combined from JASPAR (version 2022), Hocomoco and Kellis ENCODE motif database. We removed flanking bases of the TF motifs with an entropy higher than 1.9, since we observed that TF motifs with flanking bases that exhibit a high entropy have a negative effect on the fit of the distribution to the observed D_{max}. To apply our method, we pre-computed the scale parameter b for each TF motif. To do so, we randomly sampled $200,000$ SNPs from the dbSNP database (build id 154). For each TF, we computed D_{max} for all SNPs. These values are plugged in the MLE of the PDF of L_{max} and numerically solved using Newton's method to approximate b (done with the Python library scipy.optimize (Virtanen *et al.*, 2020)).
In particular, we are interested in describing the tail of the distribution of D_{max} as accurately as possible. Thus, we optimized our estimated scale parameter b by minimizing the mean squared error (MSE) for the tail of the distribution of D_{max} (25% of all values). Therefore, we decreased / increased the estimated scale parameter b by 0.01 as long as the MSE is decreased.

5.3 Collections of experimentally validated TF-SNP pairs

This section describes the datasets of validated TF-SNP pairs we utilized to evaluate our method and compare it to the performance of *atSNP*.

Allele specific binding events

We collected 1,760 Allele Specific Binding (ASB) events identified in the human cell line GM12878 using 14 TF ChIP-seq datasets (Shi *et al.*, 2016). For heterozygous binding sites of a TF, an ASB event is defined if the number of mapped ChIP-seq reads for one allele is significantly higher than for the other allele. The authors noticed that only 19.3% of the ASB events overlap with a TFBS of the TF that the ChIP-seq experiment was designed for. Hence, we only want to consider the SNPs overlapping with a TFBS of the used TF motif. Therefore, we gathered the 14 TF motifs from the JASPAR database (version 2022) (Castro-Mondragon *et al.*, 2021) and computed per TF the TFBS using FIMO (see Section 2.2.3) (version meme-5.2.0, default p-value cutoff) (Grant *et al.*, 2011). Since redundant motifs would lead to false positives later in our analysis, we clustered a combined TF motif set of the JASPAR, Hocomoco (Kulakovskiy *et al.*, 2017) and Kellis ENCODE motif (Kheradpour and Kellis, 2013) database using the similarity measure and clustering approach from Pape *et al.* (Pape *et al.*, 2008)(see Section 2.2.5). We checked which of the 14 TFs belong to the same cluster and kept per TF motif cluster only the TF with the highest number of associated SNPs. Doing so, we removed two TFs resulting in 368 SNPs for 12 TFs (see Appendix A.3.1 and A.5).

SNP-Selex data

The SNP-SELEX data was downloaded from the web portal GVAT database (Yan *et al.*, 2021). We gathered the SNPs called *original batch*. In total, Jian *et al.* studied 1,612,172 TF-SNP pairs for 271 different TFs. For each TF we collected all SNPs that have biological evidence to cause a differential binding event. Therefore, we filtered the SNPs for those which have an oligonucleotide binding score p-value < 0.05 and a preferential binding score p-value < 0.01, as proposed by the authors, resulting in 9,840 SNPs for 129 TFs. We gathered the TF motifs from Boytsov *et al.* (Boytsov *et al.*, 2022), which provide optimized PWM motifs for the SNP-SELEX dataset. As for the ASB dataset we excluded for each TF the SNPs without a TFBS for at least one of the two alleles. If less than 5% of the SNPs associated to a TF do have a TFBS, we excluded all TF-SNPs pairs for the analysis. Next we checked which of the TFs belong to the same TF motif cluster based on the clustering used for the ASB SNPs. For

each TF motif cluster we selected the TF with most SNPs, resulting in a total of 33 TFs and 1,494 SNPs (see Appendix A.3.1 and A.5).

Method performance evaluation

We applied our method and *atSNP* to the SNP-TF pairs collected for the ASB events (see Section 5.3) and SNP-SELEX data (see Section 5.3). In Appendix A.3.1 the commands used are listed. To evaluate the performance of each method, we computed a precision-recall curve and the area under the precision recall curve (AUCPR) using the R package *PRROC* (Grau *et al.*, 2015). To test whether the area under the receiver operating characteristic (ROC) curve is significantly different between two approaches we applied the method from DeLong *et al.* (DeLong *et al.*, 1988) using the R package pROC (Robin *et al.*, 2011).
We compared the run times of our approach and *atSNP* for 100, 500, 1,000, 10,000, 20,000 and 40,000 SNPs randomly sampled from the dbSNP database (build id 154) (Sherry, 2001). As both methods provide a parallel mode, we compared the runtime for 1, 8 and 16 threads (see Figure 5.3C).

5.4 Evaluation of D and D_{max} on randomly sampled SNPs

Next, we explored how well the theoretically derived Laplace distribution approximates the differential TF binding score D on randomly sampled SNPs and how the scale parameter b is chosen. Then, we investigated in how well the maximal differential TF binding score distribution can be described by the modified Laplace distribution.

5.4.1 The differential TF binding score D follows approximately the $L(0, b)$ distribution

Consider the sequences S^1 and S^2 each holding an allelic variant of a given SNP and a TF model M that characterizes the binding behavior of a TF. We defined the differential TF binding score (D) between S^1 and S^2 as the log-ratio of the p-values of the TF binding scores (see Section 5.2.1).
To investigate the distribution of the differential TF binding scores, we decided to represent the TF models with PWMs, which are widely used and easy accessible for hundreds of human TFs, and for which other methods exist for comparison. We computed D for 200,000 SNPs randomly sampled from the dbSNP database for PWMs of different length. In Figure 5.1 the resulting distributions of the differential TF binding scores for three TFs are visualized.

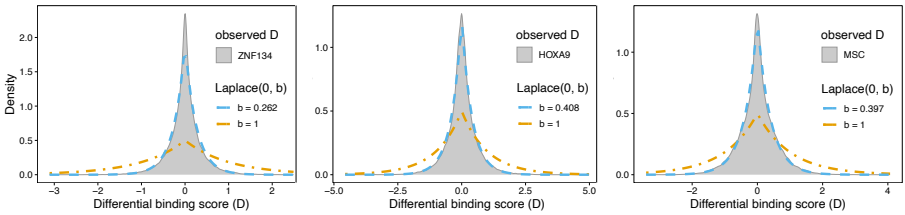


Figure 5.1: Differential TF binding score distributions for three TFs of different length. The distributions of D for the PWMs of TFs ZNF134 (length 22), HOXA9 (length 8) and MSC (length 10) are compared to the $L(0, 1)$ (orange) and $L(0, b)$ distribution (blue). The scale parameter b is estimated for each TF model separately.

For comparison, the $L(0, 1)$ distribution is displayed as well. Even though we argued that theoretically D should be $L(0, 1)$ distributed, we observed that our experiments suggest otherwise. The sequences of S^1 and S^2 just differ by one letter at the position of the SNP. Hence, the TF binding scores for S^1 and S^2 do not change much, especially if the SNP does not affect the binding site. As a result, we concluded that the corresponding p-values may not always be independent of each other. Consequently, we observed that D is close to 0 more often than one would expect for independent p-values. However, we empirically observed that D can be approximated by a $L(0, b)$ distribution with a scale parameter b fitted for each TF model M (see Figure 5.1 blue curves).

5.4.2 The maximal differential TF binding score D_{max} distribution differs between multiple TFs

In the previous section, we assumed that we are interested in D for a specific sequence position, thus the sequence S and the TF model had the same length. However, that is an unrealistic assumption. Usually, the sequence is longer than the TF model. As explained in Section 5.2.2, we computed D for all subsequences overlapping the SNP to identify where the binding affinity of a TF is most affected and keep the maximal absolute differential TF binding score (D_{max}).

For $200,000$ randomly sampled SNPs, we determined D_{max} and visualized the resulting distributions for the TFs ZNF134, MSC and HOXA9 in Fig. 5.2A. Since the distributions are different, it is not possible to compare D_{max} between different TFs directly. However, the usual application of such a statistic is to evaluate and compare the effect for several hundred TFs on a SNP set. To allow this comparison, we aimed to compute a p-value for D_{max}. We first tried to fit known distributions to the observed D_{max} values (using R package gamlss (Rigby and Stasinopoulos, 2005)), but this approach did not result

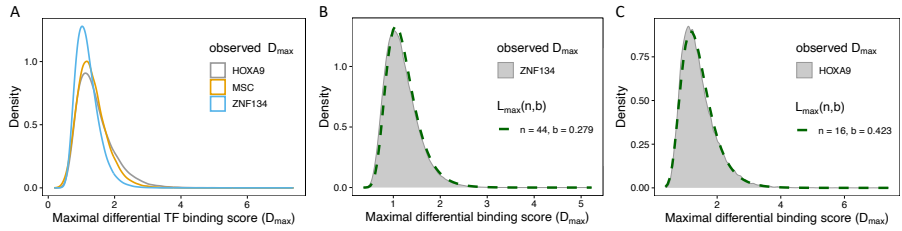


Figure 5.2: Distribution of D_{max} values. (**A**) Density plot for observed D_{max} values for the PWMs of TFs HOXA9, MSC and ZNF134 indicated by different colours. Distributions of observed D_{max} values for ZNF134 (**B**) and HOXA9 (**C**) in comparison to the $L_{max}(n, b)$ distribution, parameter values for n and b (fitted) shown in the plot.

in the same distribution for multiple TFs. As an alternative, we derived the $L_{max}(n, b)$ distribution. The parameter n denotes the length of the TF model, which is given by the number of sequence windows tested by the model and the scale parameter b is estimated for each TF separately using the MLE from Eq. (5.9).
In Fig. 5.2B, C we show example distributions of observed D_{max} values for the TFs HOXA9 and ZNF134 and the corresponding $L_{max}(n, b)$ distribution. One can observe that the $L_{max}(n, b)$ distribution accurately approximates the observed D_{max} values for each TF. Using the CDF of the $L_{max}(n, b)$ distribution we can compute a p-value for D_{max} values, which are then comparable between multiple TFs.

5.5 Evaluation on experimentally validated TF-SNP pairs

To analyze the performance of our approach, we collected TF-SNP pairs from data sources with experimental evidence that the TF is affected by the SNP. We gathered ASB events, which are defined using TF ChIP-seq data and data collected from SNP-SELEX experiments, an *in vitro* measurement of the TF-DNA interaction strength for each allele of the SNP (see Section 5.3 and 5.3). To evaluate how well a method can distinguish experimentally validated TF-SNP pairs from those that are not validated, we defined a classification task, where the positive class contains the collected TF-SNP pairs. The negative class is defined as all possible combinations of considered TFs and SNPs, excluding those from the positive class. The ASB dataset consists of 368 positively labeled TF-SNP pairs and $4,036$ negatives, and the SNP-SELEX data of $1,814$ positives and $58,162$ negatives.
Using these datasets we not only want to evaluate our approach, but also

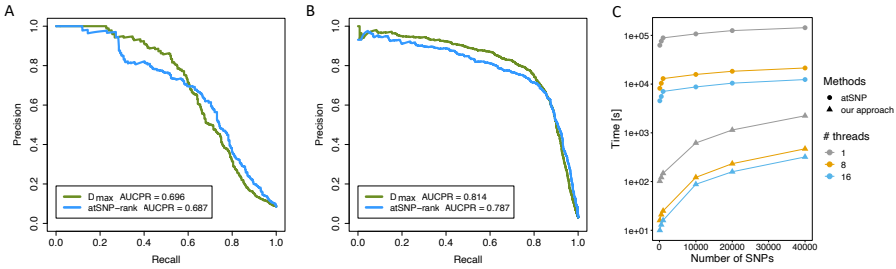


Figure 5.3: Comparison between our approach and *atSNP*. Precision-Recall curve for the ASB events (**A**) and the SNP-SELEX dataset (**B**) for the D_{max} p-value and the rank-based p-value of *atSNP*. (**C**) Runtime analysis of our approach and *atSNP*. The lineplot shows the runtime for both methods (y-axis, log10-scale) for randomly sampled SNP sets of different size (x-axis) on a different number of threads.

compare our method to the previously published method *atSNP* (Zuo *et al.*, 2015). To the best of our knowledge, *atSNP* is the fastest PWM based method currently available, which also provides a p-value for the differential TF binding score. *is-rSNP* is too slow to be applied on the large datasets considered here (Macintyre *et al.*, 2010). We applied our approach and *atSNP* for each of the two datasets separately, and evaluated the performance for the D_{max} p-value in comparison to *atSNP*'s rank-based p-value, which is recommended by the authors. The rank-based p-value indicates if the log-odds ratio of the p-values of the TF binding score for the wildtype and the alternative allele significantly differ from what one would expect by chance. Additionally, they provided a diff-based p-value, that evaluates the changes in the TF binding score directly. However, the authors of *atSNP* mention that the diff-based p-value is not reliable, and our experiments confirmed a poor performance (data not shown).

Figure 5.3A and B show the resulting precision recall curves and the AUCPR for the ASB events (Figure 5.3A) and the SNP-SELEX data (Figure 5.3B). Even though the negatively labeled TF-SNP sets are several times larger than the positive sets, a reasonable AUCPR is reached for both datasets indicating the quality of our method and the rank-based p-value of *atSNP*. The AUCPR of the D_{max} p-value is improved in comparison to the rank-based p-value of *atSNP*. For the ASB events the AUCPR of the D_{max} p-value is 0.9% higher than the rank-based p-value. For the SNP-SELEX dataset the improvement of the D_{max} p-value is 2.7% in comparison to the rank-based p-value. Additionally, for both datasets the difference of the area under the ROC curves between the D_{max} p-value compared to the rank-based p-value are significant (p-value ≤ 0.05, ASB data: p-value $= 0.00541$, SNP-SELEX data: p-value $9.457e^{-5}$)

according to the method of DeLong *et al.* (see Section 5.3).
We compared the runtime of our approach and *atSNP* for 6 randomly sampled SNP sets of sizes between 100 and 40,000 SNPs for 817 TF motifs using 1, 8 and 16 threads. As shown in Figure 5.3C our method is between 623 (500 SNPs on 1 thread) and 38 (40,000 SNPs on 16 threads) times faster than *atSNP*.
To decide for a reasonable D_{max} p-value cutoff, we computed the $F1$ score for the ASB and SNP-SELEX data. Based on the result we recommend to utilize a D_{max} p-value cutoff between 0.01 and 0.001. We combined our statistical approach in an easy-to-apply workflow with cell type-specific epigenetics data outlined in Chapter 6, where we also present typical applications.

5.6 Discussion

Throughout this chapter, we presented a new statistical approach to identify regulatory SNPs modifying the binding sites of TFs. We aimed to provide a method that allows to compute statistical significance for general TF models. We compared our new approach to the previous method *atSNP* in terms of performance and runtime and demonstrate that our new approach is as least as accurate as *atSNP*. By comparing the runtimes of both methods, we showed that our approach is extremely fast also for large sets of SNPs and hundreds of TFs.
To test our approach, we used PWMs as TF model, since they are still commonly used and available for hundreds of human TFs. Nevertheless, our approach can be applied to any other TF model once a p-value for the TF binding score is computed, which can be done using Monte Carlo sampling if not otherwise existing. Thus, an interesting research direction is to explore the $L_{max}(n, b)$ distribution fit to observed D_{max} values of other approaches, such as TFFMs or SLIM models.
Our approach does not directly take into account cell type or tissue specific information. However, a useful approach is to exclude motifs from TFs not expressed in the cell type or tissue of interest to reduce the number of false positive predicted rSNPs. Further, one can easily combine the predicted rSNPs with other epigenomic data as outlined in the next Chapter.
Additionally, we want to emphasize that we combined several TFs within one dataset to compare the methods, resulting in a highly imbalanced dataset. In our opinion, this is a more realistic evaluation setting than evaluating the TFs one by one as is often done.
Another advantage of our approach is that it has no significant additional runtime compared to widespread score-based approaches that do not assess significance. One only needs to precompute the scale parameter b for the used TF motifs. On our GitHub repository (https://github.com/SchulzLab/SNEEP

we provide our approach implemented in C++ and the precomputed scales for the 817 TF motifs used for the presented analyses.
We believe that our approach will be helpful to identify novel rSNPs and thereby contribute to the understanding of molecular mechanisms leading to various traits and diseases.

Chapter 6

SNEEP: SNP exploration and functional analysis using epigenetic data

In this chapter, we outline the Sneep workflow that combines our statistical approach to detect rSNPs and the affected TFs (see Chapter 5) with cell type-specific epigenomics data. Sneep is the abbreviation for SNP exploration and analysis using epigenomics data. In the second part of this chapter, we provide two applications on how one can use Sneep to detect TFs highly affected by rSNPs as well as to identify disease associated genes based on rSNPs.

6.1 Incorporation of epigenetic data to our approach detecting rSNPs

Depending on the type of application we combined our statistical approach (see Chapter 5) with different functionalities, resulting in our Sneep workflow. The workflow itself is not published yet, but we provide the source code and a detailed description on our github repository (https://github.com/SchulzLab/SNEEP). Our bioconda package from the previous chapter includes also the functionalities of Sneep. In the following we list the contribution of the people involved:

- Implementation, testing and visualization: myself
- Design of the workflow: Marcel H. Schulz and myself
- Set up bioconda package: Laura Rumpf

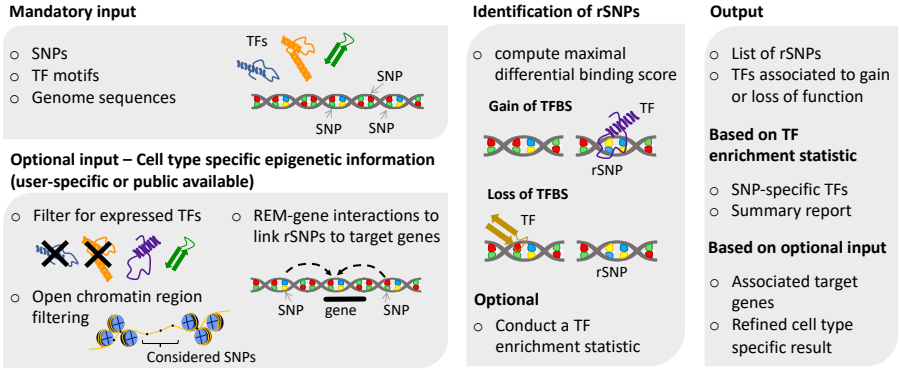


Figure 6.1: Overview of the *SNEEP* workflow. Expected and optional input to our SNEEP workflow is visualized (left panel). The middle panel illustrates how to detect rSNPs with our statistical approach, their effects and the optional possibility to perform a TF enrichment statistic. Based on the used parameters SNEEP provides different kinds of outputs (right panel).

6.1.1 The SNEEP workflow

SNEEP is a computational workflow to assess the differential TF binding effects of non-coding SNPs (see Figure 6.1). Our workflow expects three mandatory inputs: (1) A list of candidate SNPs, (2) a set of TF motifs given as count matrices in TRANSFAC format and (3) the genomic sequence of a species of interest. Using our statistical approach presented in the previous chapter, we detect rSNPs and the corresponding affected TFs efficiently for thousands of SNPs and large sets of motifs. Additionally, SNEEP identifies TFs which are statistically significant more often associated to a gain or a loss of function than expected. Further, we provide the optional functionality to identify TFs specific to the input SNPs by affecting TFBS more often than one would expect on random SNP lists. Therefore, we computed a TF enrichment statistic, which can be done in a reasonable amount of time even for large SNP sets because of the efficiency of our statistical approach. Three optional inputs can be provided to include cell type or tissue specificity, which results in a more accurate prediction of rSNPs. SNEEP can incorporate (1) gene expression data to exclude TFs not expressed in the cell type or tissue of interest, (2) DNase1-seq or ATAC-seq data to filter for SNPs in open chromatin and (3) REM-gene interactions to link rSNPs to target genes. One can combine the functionalities of our workflow in all possible ways which provides a high flexibility to the user. In addition to a plain text file, we provide as output an automatically generated summary report. In the following, we provide more details of our SNEEP workflow.

6.1.2 Identification of regulatory SNPs

To identify rSNPs we developed a statistical approach that computes the maximal differential binding score D_{max} based on the log-odds ratios of the binding scores of the two allelic variants of a SNP. D_{max} can be well approximated by a modified Laplace distribution, which allows us to derive a p-value. Hence, we can compare the effect of a SNP on multiple TFs. Our statistical approach is explained in detail in Chapter 5. The Sneep workflow is using PWMs as TF models. Instead of using our provided collection of 817 human TF motifs gathered from the JASPAR, Hocomoco and Kellis ENCODE motif database, it is possible to use a customized TF motif set, also for species other than human.

6.1.3 Identification of TFs associated to a gain or a loss of function

Given a SNP with significant maximal differential binding score D_{max}, we define a loss of function when the wild type allele has a higher binding score than the alternative allele. Similarly, the gain of function is given when the binding score of the wild type allele is lower than the binding score of the alternative allele. We evaluate whether TFs over all considered SNPs are more often associated to a gain or loss of function. Therefore, we compute per TF a p-value using a two sided binomial test, assuming that gain and loss of a TF binding site occur equally likely ($p = 0.5$). The TFs associated to a gain or loss of functions (p-value ≤ 0.1) are listed in our summary report.

6.1.4 Incorporation of cell type-specific epigenetic information and REM-gene links

With our Sneep workflow, we want to offer the user a flexible way to include customized cell type or tissue specific epigenetic information and thereby improve the reliability of our prediction. Therefore, one can provide RNA-seq data in form of TPM or FPKM values to remove motifs from the given motif set of TFs not or only weakly expressed. Further, it is possible to provide cell type or tissue specific regions of interest, like open chromatin data. Sneep excludes all given SNPs not overlapping with these regions. Both options enable us to refine the results of our statistical approach to a cell type or tissue specific prediction.
In addition, REMs linked to their target genes are used to associate rSNPs overlapping with these REMs to potential target genes. Sneep provides a collection of REM-gene interactions incorporating $2,404,861$ REMs downloaded from the EpiRegio database (see Chapter 4), $57,538$ promoter regions of annotated genes (gencode, release 38) and $60,415$ interactions identified with the ABC score on heart data (Anene-Nzelu *et al.*, 2020; Fulco *et al.*, 2019). It is

also possible to combine customized REM-gene interactions with our collection or to provide them solely to Sneep.

6.1.5 Identification of SNP-specific TFs using a TF enrichment statistic

In order to identify TFs that are more often affected by the given SNPs than expected for random SNPs, we implemented a TF enrichment statistic. This is necessary because each TF motif occurs with a certain probability in the genomic sequence depending on the properties of the TF motif itself. Thus, one can not directly compare the counts how often a TF was affected. To perform the TF enrichment statistic, we randomly sampled at least 100 SNP sets from the dbSNP database (build id 154), which contain the same number of unique SNPs as the corresponding input SNPs. If for the given SNPs the minor allele frequency (MAF) (see Section 2.1.9) is provided, the MAF distribution is kept the same for the randomly sampled SNP set. Thus, we ensure the same distribution of rare and common SNPs for the input data and the background dataset. Using our statistical approach, we compute D_{max} separately for each SNP within the given SNP list and the randomly sampled SNP sets. Next, we counted per TF how often the binding sites were significantly affected (D_{max} p-value ≤ 0.01). We denote this count per TF for the given SNPs as $TFcount$ and we took the mean count over all randomly sampled SNP sets, denoted as $bgCount$. To identify TFs that are more often affected by the given SNPs than expected, we computed an odds-ratio for each TF as:

$$\text{odds-ratio(TF)} = \frac{\alpha \ / \ (1-\alpha)}{\beta \ / \ (1-\beta)}, \tag{6.1}$$

where $\alpha = \frac{TFcount}{\#SNPs}$, $\beta = \frac{bgCount}{\#SNPs}$ and $\#SNPs$ is the number of unique SNPs in the input SNP list. TFs that occur twice as often as expected (odds-ratio > 2) are assumed to be specific for the given SNPs and are listed in the summary report.

6.1.6 Automatic generation of a summary report

We combined the results of the Sneep workflow in an automatically generated summary report done with RMarkdown that contains:

- a general overview of the analysis and used parameters/ input files
- TFs associated to a gain or loss of function,
- TFs specific to the given SNPs and the corresponding odds-ratio (if TF enrichment statistic was performed),
- the target genes linked to the rSNPs,

- a functional annotation analysis done using g:Profiler (Raudvere *et al.*, 2019) based on the associated target genes.

Appendix Section A.4.1 outlines parts of the summary report for the GWAS atherosclerosis.

6.2 Application: Identification of TFs that are altered by genetic variants mediating gene expression

We applied our Sneep workflow to large eQTL studies on the cell types fibroblasts and lymphocytes aiming to identify cell type-specific TFs. This application is also published in our preprint *A statistical approach to identify regulatory DNA variations* (Baumgarten *et al.*, 2023), thus most of the text and figures are similar to the publication. The contribution for this application is the following:

- Computational steps, generating plots and writing the text: myself
- Choosing a suitable dataset, discussion, evaluation and literature search: Marcel H. Schulz and myself
- Proofreading: all authors

6.2.1 Detection of cell type-specific TFs affected by eQTLs in fibroblasts and lymphocytes

Identifying cell type-specific regulators with modified binding behavior induced by genetic variants associated to genes, might be helpful to unravel regulatory pathways or molecular mechanisms. We aim to identify TFs more often affected by a set of eQTLs than one would expect on random data. As an example, we applied our Sneep workflow including the TF enrichment statistic to $14,722$ eQTLs associated to lymphocytes and $45,917$ eQTLs associated to fibroblasts gathered from the GTEx portal (details are provided in Appendix Section A.4.2).
For each eQTL, the D_{max} p-value for 817 human TFs was computed and for a reliable TF enrichment statistic $1,000$ randomly sampled SNP sets were obtained. Within the TF enrichment statistic the odds-ratio from Equation (6.1) between the number of binding sites affected by a TF over all eQTLs and their average number of affected binding sites on the sampled SNP sets is computed (see Section 6.1.5), to identify TFs more often affected by the eQTLs than expected.
For fibroblasts 57 TFs and for lymphocytes 66 TFs were detected (full list per cell type in Appendix Section A.4.2. Figure 6.2A visualizes the odds-ratios of

the analyzed TFs of fibroblasts against lymphocytes, the enriched TFs with an absolute difference of the odds-ratio > 1 are labeled. For several of the cell type-specific TFs, we found evidence in the literature: For instance, it has been shown that in skin fibroblasts EGR3 can upregulate genes associated with tissue remodeling and wound healing (Fang *et al.*, 2013). The expression of TF SNAI1 in cancer associated fibroblasts is directly associated to chemoresistance via the mediation of the extracellular matrix (Galindo-Pumariño *et al.*, 2022). For lymphocytes, we identified several highly expressed TFs from the Ets-related TF family, among others, with additional evidence from the literature. For instance, in mice it has been shown that the TFs ELK1 and ELK4 function redundantly to restrict the generation of innate-like CD8+ T-cells (Maurice *et al.*, 2018). Recently, Tsiomita *et al.* showed that ERF is a potential regulator during T-lymphocyte maturation (Tsiomita *et al.*, 2022). Further, for innate lymphoid cells (ILC), which are a population of lymphocytes, it can be shown that ELK3 is regulated by the circRNA circTmem241. The knockdown of ELK3 significantly decreased the number of ILCs (Liu *et al.*, 2022).

6.2.2 Evaluation of the difference in gene expression between cell type-specific and non cell type-specific TFs

We cannot provide literature evidence for all cell type-specific TFs, therefore we want to evaluate the expression values of enriched TFs in comparison to TFs not enriched (Figure 6.2B). Motifs from the same TF family have often similar binding preferences, which results also in similar motifs. To avoid motif redundancy, we clustered the TF motifs (see Section 2.2.5). If a TF within a cluster occurs twice as often as expected, we consider it as enriched. Maximal TPM value over all TFs in the cluster is used to evaluate the gene expression level. According to a one-sided Wilcoxon rank sum test the differences in gene expression between the two groups are seen significant (fibroblasts: p-value 0.0016, lymphocytes p-value 0.028).

6.2.3 Conclusion

Using our SNEEP workflow, we identified eQTLs affecting a TF binding site and pointed to cell type-specific TFs by performing a TF enrichment statistic. This analysis is easily possible even for a large number of randomly sampled SNP sets, since the runtime of our statistical approach behind SNEEP is extremely fast. A literature search and the significant differences in gene expression between cell type-specific and non cell type-specific TFs supports that many of the identified TFs are likely to mediate gene expression differences and are interesting candidates for further cell type-specific investigations.

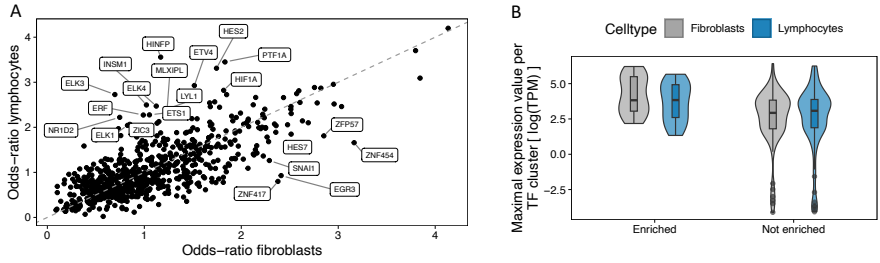
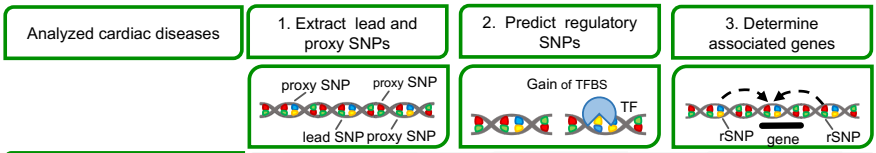


Figure 6.2: Identification of TFs altered by eQTLs in lymphocytes and fibroblasts. (**A**) Scatter plot showing the computed odds-ratio of a TF for eQTLs from fibroblast cells (x-axis) against lymphocyte eQTLs (y-axis). The TFs are labeled if the odds-ratio > 2 and the absolute difference of the odds-ratio between the two cell types is > 1. (**B**) Violin-boxplot illustrating the differences in terms of expression values between the TFs which are more often affected by the eQTLs (enriched odds-ratio) than expected (not enriched odds-ratio) for two different cell types (coloring).

6.3 Application: Identification of cardiovascular disease associated genes using regulatory SNPs and REM-gene interactions

In the following, we summarize the bioinformatic analyses of our work *CVD-associated SNPs with regulatory potential drive pathologic non-coding RNA expression*, which is available as preprint on bioRxiv (Zhu *et al.*, 2023). Parts of the figures and the text below are taken from the preprint. The work is a collaboration with the lab of Jaya Krishnan. Chaonan Zhu and myself share the first authorship. We list the contributions of all authors in the following:

- Conduction of the bioinfomatics analyses and generation of the corresponding figures: myself
- Discussion, planning and evaluation of the bioinformatics analyses and their results: Marcel H. Schulz (M.H.S) and myself
- Planning and design of the experiments, Chaonan Zhu (C.Z), Ting Yuan (T.Y), and Jaya Krishnan (J.K)
- Experimental validation: C.Z with help of Meiqian Wu, Yue Wang, Arka Provo Das, Maria Duda and Stefanie Dimmeler
- Providing cells: Jaskiran Kaur, Thanh Thuy Duong and Minh Duc Pham
- Co-expression analysis on GTEx data: Fatemeh Behjati Ardakani
- Paper writing C.Z, T.Y, M.H.S, J.K and myself
- Proof reading: all authors



Analyzed cardiac diseases	1. Extract lead and proxy SNPs	2. Predict regulatory SNPs	3. Determine associated genes
Coronary artery disease	10820	954	652
Cardiac arrhythmia	6709	491	358
Myocardial infarction	1204	95	56
Cardiomyopathy	2637	314	221
Myocardial ischemia	277	21	14
Aortic stenosis	440	30	8

Figure 6.3: Overview of the studied cardiovascular diseases, identified rSNPs and associated target genes. We collected SNPs for 6 different cardiovascular diseases from the NHGRI-EBI GWAS Catalog. With our SNEEP workflow, we identified regulatory SNPs, that are predicted to have an impact on TFBS. Using the optional functionality to provide REM-gene interactions, the rSNPs were linked to putative target genes. For each step the number of SNPs or genes for each disease is given per row.

6.3.1 Detection of rSNPs and their target genes in cardiovascular diseases GWAS

The aim of this study was to identify non-coding RNAs (ncRNAs) associated to cardiovascular diseases. A possible strategy to detect such ncRNAs is to apply our SNEEP workflow on GWAS of cardiavascular diseases using the optional functionality to integrate REM-gene interactions to link rSNPs to potential target genes. Other alternative strategies how to link SNPs to genes are outlined in Section 2.2.8.

We retrieved significant lead SNPs from the EBI GWAS Catalog (Buniello *et al.*, 2018) for the cardiovascular diseases *Coronary artery disease*, *Cardio arrhythmia*, *Myocardial infarction*, *Cardiomyopathy*, *Myocardial ischemia* and *Aortic stenosis* (see Appendix Section A.4.3 for more details). Next, we collected associated proxy SNPs ($R^2 > 0.75$) and applied our SNEEP workflow separately to each of the resulting SNP sets. We used over 2.4 million regulatory elements from the EPIREGIO database as REM-gene interactions (details on the used commands can be found in Appendix Section A.4.4). Figure 6.3 shows for each disease the number of lead and proxy SNPs, the number of predicted rSNPs and the number of associated genes.

Applying our SNEEP workflow to cardiovascular disease GWAS identified hundreds of protein coding and non-coding genes (see Figure 6.4A) Interestingly, the amount of protein coding and non-coding genes is often similar. To sys-

tematically analyze the properties of the associated genes, we visualized the number of REMs affected by at least one rSNP per gene against the number of rSNPs linked to a gene (Figure 6.4B). We noticed that protein coding and non-coding genes behave similarly in the number of associated rSNPs and also in the number of affected REMs. We speculate that candidates for further investigation might be genes with either a high number of REMs with affected TFBS (upper right corner) or genes with REMs affected by more than one rSNP (genes above the diagonal). Using these criteria and the expertise of Jaya Krishnan's group, we selected a few interesting non-coding candidate genes. However, before validating these genes in the lab, we investigated if the associated genes are meaningful in the context of the studied diseases.

6.3.2 Disease enrichment analysis supports that the predicted target genes might be involved in cardiovascular diseases

As a proof of concept, we first checked if the associated protein-coding genes are known to play a role in the underlying cardiovascular diseases. Therefore, we performed a diseases enrichment analysis using the DisGeNET database, a large collection of known disease associated genes. The analysis computes whether the identified protein-coding genes are enriched among the previously associated disease genes within DisGeNET. For 4 of the studied diseases we were able to run the analysis and detect significantly enriched phenotypes related to the disease (see Appendix Section A.4.6). For *Myocardial ischemia* and *Aortic stenosis* less than 30 protein coding genes were predicted. Thus, we omitted the disease enrichment analysis for these diseases because it would be statistically underpowered.
Since the DisGeNET database mostly contains protein coding genes, we cannot directly apply the same analysis for the non-coding genes. However, since we were focused on ncRNA in this work, we aimed to provide evidence that also the non-coding genes are relevant for the studied diseases. First, we searched in the literature and existing databases, such as the Heart Failure database for known RNA biomarkers (HFBD) (He *et al.*, 2021) if the identified non-coding genes are associated to the studied diseases. But none of our associated non-coding genes was listed in existing databases. As an alternative we used an indirect strategy to test whether the non-coding genes of a cardiovascular disease were significantly associated to the corresponding disease. By assuming that non-coding genes might have a direct or indirect effect on co-expressed genes, we conduct a co-expression analysis as illustrated in Figure 6.5A. Therefore, we downloaded $9,662$ RNA-seq samples from various cell types from the GTEx portal (GTEx Consortium, 2017). By using Spearmans correlation coefficient as similarity measure, we compared the expression of protein-coding genes with the expression of the identified non-coding genes. We gathered the top 10 most correlated

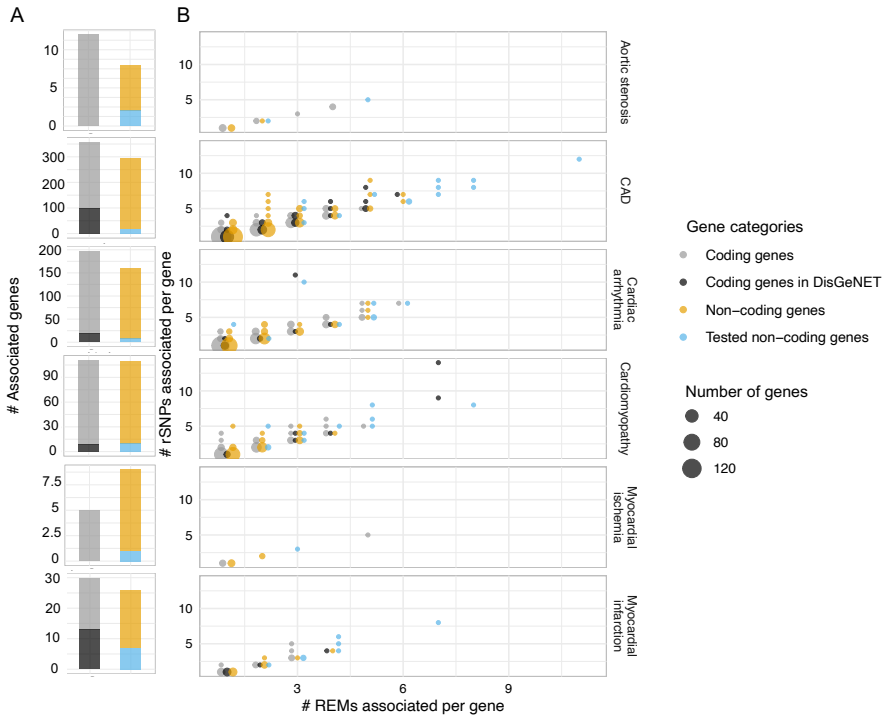


Figure 6.4: Analysis of identified genes associated to rSNPs separated in protein coding and non-coding genes. (**A**) A stacked bar plot per GWASs shows the number of associated genes per category. (**B**) Dot plot showing, per gene, the number of REMs affected by at least one rSNP (x-axis) and the number of rSNPs overall REMs linked to a gene (y-axis) separated per GWAS. Genes are grouped into protein-coding, protein-coding associated with the disease according to DisGeNET (see Appendix Section A.4.5), non-coding, and non-coding genes experimentally studied in our paper (Zhu *et al.*, 2023) (circle colour). The dot size correlates with the number of genes having the same x- and y- coordinate values.

protein coding genes for each non-coding gene (see Figure 6.5B). The resulting set of protein coding genes for each disease is highly co-expressed to the non-coding genes identified with our SNEEP pipeline. These co-expressed protein coding genes were used to assess a disease enrichment analysis based on the DisGeNET database. We list the used commands in Appendix Section A.4.6. For the cardiovascular diseases with more than 30 non-coding genes, we observed a strong enrichment of disease related phenotypes, as shown in Figure 6.5C.

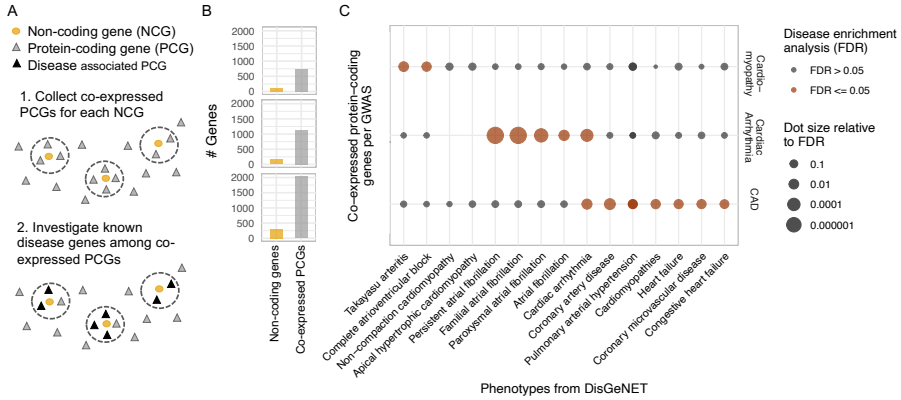


Figure 6.5: Disease enrichment analysis of protein coding genes co-expressed with non-coding genes associated to the studied cardiovascular diseases. (A) Graphical illustration of the co-expression analysis, where for each non-coding gene (NCG, circle) highly co-expressed protein-coding genes (PCGs, triangles) are identified. Using the strongest correlated protein-coding genes, a disease enrichment analysis based on the DisGeNET database is computed. (**B**) Barplot visualizing the number of non-coding and co-expressed protein-coding genes per cardiovascular disease (row). (**C**) Dot plot showing enriched cardiovascular phenotypes from the DisGeNET database (x-axis) for the top 10 co-expressed protein-coding genes for the studied cardiovascular disease with more than 30 associated non-coding genes (y-axis). The dot coloring represents whether a phenotype is significantly enriched (FDR<= 0.05) and the dot size is relative to the FDR.

As a result, we concluded that the majority of the identified non-coding genes might be relevant in the underlying cardiovascular disease, also including the 40 ncRNA selected for validation.

6.3.3 *IGBP1P1* affects cardiac function in an *in vitro* model

Motivated by these findings, Jaya Krishnan's lab studied 40 promising non-coding candidate genes (blue marked in Figure 6.4B) in *in vitro* models mimicking the human heart for normal and disease conditions. They found that ncRNA *IGBP1P1* regulates cell size (Figure 6.6A) and cardiomyocyte contractility (Figure 6.6B-C) and concluded that inhibiting *IGBP1P1* may improve cardiac function in cardiovascular diseases. A detailed description of their experiments, can be found in our preprint *CVD-associated SNPs with regulatory potential drive pathologic non-coding RNA expression* (Zhu *et al.*, 2023).

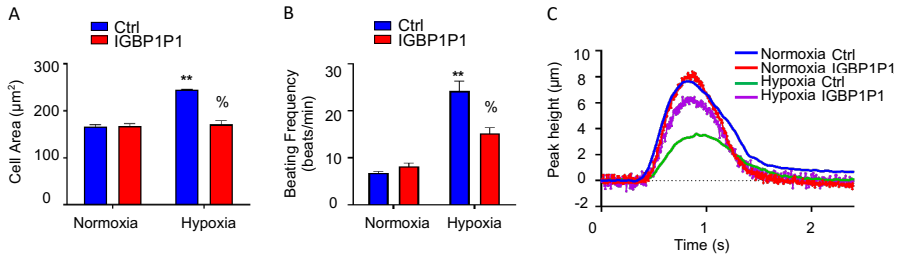


Figure 6.6: Results of the experimental validation of the ncRNA IGBP1P1. Barplots show the difference between control and *IGBP1P1*-inhibited human cardiac organoids for cell size (**A**) and beating frequency (**B**) for normal (normoxia) and disease (hypoxia) conditions. The differences between the conditions are significant, where p-values less or equal 0.05 and 0.01 are denoted with $\%$ and $**$, respectively. (**C**) Tracks of the beating frequency of control and *IGBP1P1*-inhibited human cardiac organoids for normal and disease conditions are visualized.

6.3.4 Conclusion

In the presented application, we used our Sneep workflow to identify genes that might be affected by SNPs associated to a specific disease. Therefore, we detected the rSNPs and linked them to their target genes. Further, we applied a disease enrichment analysis using DisGeNET that supported our findings. Experimentally, Jaya Krishnans lab verified that *IGBP1P1* affects cardiac function in an *in vitro* model mimicking the human heart.
To sum up, we demonstrated that our Sneep workflow combined with REM-gene interactions can be extremely helpful to pinpoint to candidate genes worth investigating.

6.4 Discussion

In this chapter we introduced our computational Sneep workflow to predict regulatory SNPs (rSNPs) and combined it with customized cell type or tissue specific epigenetic data. For example, it is possible to provide RNA-seq data of the studied cell type or tissue to exclude TFs not expressed, which reduces the number of false positively predicted TF-SNP pairs. Further, one can limit the analysis to SNPs only in accessible regions of the genome, if open chromatin data is given. We also allow to link rSNPs to genes by using publicly available REM-gene interactions, such as provided from EpiRegio. Additionally, we implemented functionalities which allow to interpret the result in several different ways. We highlight TFs which are associated to a loss or gain of function and TFs that are more often affected than expected according to random SNP

sets. The results of our Sneep workflow are visualized in a summary report specific for the given input SNP data. Within this summary report, we also provide a table holding the associated genes and the number of affected rSNPs. Thus, one can identify the genes whose regulation is affected by several rSNPs, similar as in Figure 6.4.
We showed two possible applications of our Sneep workflow. In the first one, eQTLs from fibroblasts and lymphocytes were used to point to cell type-specific TFs. To do so, we applied our TF enrichment statistic to compute a reliable background distribution based on randomly sampled SNP sets. This is easily possible because of the speed of our approach. We noticed that this kind of analysis is rarely done, however, it can be useful to identify TFs which might be interesting candidates for further investigation. For the outlined example, we found for several of the predicted cell type-specific TFs evidence in the literature that associates it to the studied cell types. Further, the cell type-specific TFs had higher average expression values than the non cell type-specific TFs, which additionally supported their importance (see Figure 2.13B).
In the second application, we aimed to point to genes associated to a given phenotype using the SNPs resulting from a GWAS. We were not interested in which TF is affected, we used our statistical approach to detect the SNPs which have an regulatory effect on the gene expression. Linking these rSNPs to genes allowed us to highlight genes which might have a changed gene expression. Other approaches to link SNPs to genes exist (outlined in Section 2.2.8), such as the gene-based test. Rare SNPs are commonly omitted by a gene-based test because no linkage disequilibrium information is available. Thus, especially for rare SNPs our approach is favorable. If the SNP position and the wildtype and alternative allele is known, our approach can decide whether it is an rSNP or not.
To sum up, our Sneep workflow, which combines our statistical approach to detect rSNPs with epigenetic data and additional functionalities, is flexible and fast, as well as suitable for many applications. We believe the workflow helps to identify novel rSNPs, disease associated target genes and cell type-specific TFs and by that contributes to better understand the molecular mechanisms causing various traits and diseases.

Chapter 7

Conclusion and outlook

In this thesis, we contributed to a central question in genetics, which is to better understand how genes are regulated. In the scope of this thesis three main topics were addressed.

How to precisely predict the motif that describes a TFs' binding behavior? Is it possible to improve existing methods by incorporating the DNA binding domain?
In Chapter 3, we addressed these questions by introducing our new method MASSIF. We showed that on human TF ChIP-seq data the performance of existing TF motif enrichment tools, such as *CentriMo* (Bailey and Machanick, 2012) and *PASTAA* (Roider *et al.*, 2009) do benefit from the incorporation of the TF's DNA binding domain (DBD) information. Based on the observation that TFs with the same DBD often have similar motifs, we developed a domain score. It evaluates the similarity of a given motif to the motifs of TFs that belong to the same DBD. The domain score is either used as a filter, to exclude unlikely candidate motifs before applying existing TF motif enrichment tools, or utilized in a meta analysis, where the p-value of the domain score is combined with the p-values of existing tools. To conclude, our method MASSIF can be used to link a binding motif to a TF.
A limitation of MASSIF is that TF ChIP-seq data is required to assign a motif to a TF. This is because we incorporate *CentriMo*, which is a TF enrichment tool only working on TF ChIP-seq data. For *PASTAA* it is necessary to provide regions where it is assumed that the TF binds to, which is fulfilled by TF ChIP-seq regions. However, we noticed that often for less studied TFs where no TF binding motif is known, also no ChIP-seq data is available. In those cases, our current version of MASSIF cannot be applied. Nevertheless, it might

be possible to use the concept of the domain score. To do so, we would have to exclude *CentriMo* and derive the set of input sequences for *PASTAA* with a different strategy. For instance, we could assume that genes which are co-expressed with the studied TF might also be regulated by this TF. For these co-regulated genes, we could infer REMs using for instance EpiRegio to derive genomic regions, where we assume that the TF might bind to. However, this is a more complex task, since such an input sequence set contains much more false positives than TF ChIP-seq data. Similar to our current version of Massif we could use the domain score either as filter or in a meta analysis. Adapting Massif in such a way would make the method much more flexible to apply since only expression data of the studied TF is required. Since we did not explore the performance of *PASTAA* on such less specific sequence sets, we do not know how robust it is against false positive sequences and, thus, a more thorough analysis is required.
Throughout this chapter and the entire thesis we used PWMs to represent the binding preferences of TFs. PWMs are sometimes criticized because they assume that the positions within a motif are independent of each other, which is a simplification and therefore not always true. Several other models are available to describe the binding behavior of a TF, such as SLIM (Keilwagen and Grau, 2015), BEM (Zhao *et al.*, 2012) or TFFM (Mathelier and Wasserman, 2013) (see Section 2.2.3), which try to overcome this drawback. However, we decided that PWMs are the best model for us to use, since they are available for hundreds of TFs, easy to handle and commonly applied, as for example in the TF motif enrichment tools *CentriMo* and *PASTAA*.
As already outlined in Section 3.4, the concept of Massif's domain score is independent of the TF model used. The observation that TFs with the same DBD tend to have similar motifs, should be true for any TF model. If a similarity measure for the considered TF model exists, the score can be derived.

How to detect REMs and their associated target genes? How to identify those REMs active in the studied cell type or tissue? Is such information publicly available and how to easily access it?
In Chapter 4, we dealt with these questions and introduced our EpiRegio webserver, a resource that holds REMs associated to target genes predicted by *STITCHIT* (Schmidt *et al.*, 2021). Our webserver is publicly available and allows to easily query for REMs associated to genes and REMs overlapping genomic regions. For each REM, we provide a score describing the activity of the REM in the cell type or tissue of interest in comparison to all other cell types included in the database. One can prioritize REMs based on their contribution in the studied cell type or tissue. Based on the cell type-specific REMs, we have identified an interesting candidate TF *KLF4*. In cooperation with the lab of Stefanie Dimmeler, we showed that *KLF4* regulates gene ex-

pression changes in circPLOD2 depleted pericytes. In a second application, we used EpiRegio to detect candidate genes affected by non-coding target sites of a genome-wide CRISPR-Cas screen that induce chemotherapy resistance. Further, we identified TFs enriched in the active REMs that overlap a target site of the screen. Several of the identified genes and TFs are known to play a role in cancer growth or chemoresistance.
With the EpiRegio webserver we provided a resource that can be used to answer a variety of scientific questions in many different scenarios. For instance, it is possible to explore gene regulatory networks of genes based on genomic regions, to identify target regions for experimental setups or to analyze candidate genes of TFs. However, EpiRegio is currently limited to human, even though such a resources would also be beneficial for other species, such as mouse. Given a suitably large data set, a similar webserver could be set up for arbitrary species.
The EpiRegio webserver currently contains REMs identified on 46 cell types and tissues. It could be that the studied cell type or tissue is not included in EpiRegio, as it also happened throughout this thesis. To still consider the REMs in a cell type-specific manner, we either filtered for those REMs overlapping with open chromatin regions of the studied cell type (see Section 4.2) or computed the cell type or tissue DNase1 signal similar as we did for the cell types and tissues included within our EpiRegio webserver (see Section 4.3). To avoid such additional steps, we currently run *STITCHIT* on the larger dataset provided by the IHEC consortium (Bujold *et al.*, 2016) and update our EpiRegio webserver. These efforts are going to improve the REMs, their association to genes and the coverage of cell types.

How to assess non-coding genetic variations? What is their impact on TF binding sites and how do they affect gene expression?
In Chapter 5 we explained our statistical approach to evaluate whether a SNP is regulatory. Therefore, we introduced the maximal differential TF binding score, which can be well approximated by a modified Laplace distribution. As a consequence, we can easily compute a p-value for this score. On experimentally validated TF-SNP pairs our approach improves the results in comparison to the previously established method *atSNP*, while being an order of magnitude faster. In the following Chapter 6, we combined our statistical approach with epigenetic data and further functionalities to our Sneep workflow. This workflow allows to exclude TFs not expressed in the studied cell type or tissue, to link regulatory SNPs (rSNP) to target genes using REM-gene interactions (*e.g.* provided by EpiRegio), or to conduct a TF enrichment statistic to identify TFs highly affected by the given SNPs. To outline how flexible the Sneep workflow can be used in various kinds of settings we provided two applications. We utilized our Sneep workflow to point to TFs which are at least two times

more often affected by eQTL SNPs of lymphocytes and fibroblasts than expected. Several of these TFs are known to be important for the studied cell types. In the second application, we were able to identify genes associated to cardiovascular diseases by linking rSNPs detected with SNEEP to genes using REM-gene interactions. Among them is the non-coding gene *IGBP1P1* which has been verified in the lab of Jaya Krishnan to affect cardiac function in a cardiac organoid model.

Motivated by this finding, it might be interesting to evaluate GWAS in a large scale not only for cardiac diseases. Given the speed of our statistical approach, we could for instance determine rSNPs for all GWAS listed in the GWAS Catalog (Buniello *et al.*, 2018). Further, we could link the rSNPs to target genes and perform the TF enrichment statistics. The results might contain interesting candidate TFs and genes worth investigating in more detail and therefore, this might be an inspiring resource for the community.

Tools that assess rSNPs, such as *atSNP* (Zuo *et al.*, 2015) or *is-rSNP* (Macintyre *et al.*, 2010), rely on PWMs, which are sometimes criticized for the reasons explained above. However, a recent publication of Boytsov *et al.* (2022) outlined that PWMs are powerful enough to evaluate non-coding SNPs and that more complex models, such as the support vector machine they compared to, do not automatically result in a significantly better prediction. Since our statistical approach to identify rSNPs is independent of the TF model used, we nevertheless think it would be interesting to try a different TF model, such as SLIM or TFFM instead of PWMs and compare the performance.

Further, our approach is independent of the studied cell type or tissue. Currently, we incorporate the cell type-specificity by including epigenetic data of the studied cell type and by cell type-specific REM-gene interactions. However, an interesting idea could be to combine the outcome of our statistical approach with a model that evaluates the effect of a SNP in a more cell type-specific manner. Using models like the Basenji (Kelley *et al.*, 2018) or Enformer (Avsec *et al.*, 2021b), which are deep convolutional neural networks that predict cell type-specific epigenetic profiles and thus could be used to assess the effect of a SNP on the chromatin level. Such models are usually not able to infer the affected TF. Thus, a meta analysis that combines the p-value of our statistical approach and the p-value of a more cell type-specific model might be a possibility to detect cell type-specific rSNPs.

Appendices

In this section, we provide further tables and figures, detailed information how our methods are applied and also further details of our analyses separately for each chapter.

A.1 Further information to Chapter 3

In the following we outline in more detail which DNA-binding domains are considered in our DBD collection, and explain how we prepared the ChIP-seq data for evaluating our approach.

A.1.1 DNA binding domains of the TFs within the DBD collection

Table 1 lists the 30 DBDs observed within the JASPAR database and the structural TF class according to the TFClass system. The structural TF class was used as DBD.

DBD number	DBD	# clusters
DBD 1	Basic helix-loop-helix factors (bHLH)	42
DBD 2	Basic leucine zipper factors (bZIP)	28
DBD 3	Nuclear receptors with C4 zinc fingers	22
DBD 4	Homeo domain factors	57
DBD 5	High-mobility group (HMG) domain factors	13
DBD 6	Tryptophan cluster factors	23
DBD 7	SMAD/NF-1 DNA-binding domain factors	4
DBD 8	STAT domain factors	5
DBD 9	Other C4 zinc finger-type factors	4

DBD 10	Rel homology region (RHR) factors	8
DBD 11	Heat shock factors	2
DBD 12	p53 domain factors	2
DBD 13	C2H2 zinc finger factors	35
DBD 14	C2CH THAP-type zinc finger factors	1
DBD 15	Fork head / winged helix factors	21
DBD 16	T-Box factors	10
DBD 17	MADS box factors	4
DBD 18	DM-type intertwined zinc finger factors	1
DBD 19	ARID domain factors	2
DBD 20	Heteromeric CCAAT-binding factors	1
DBD 21	Paired box factors	6
DBD 22	Basic helix-span-helix factors (bHSH)	8
DBD 23	Runt domain factors	2
DBD 24	TEA domain factors	4
DBD 25	GCM domain factors	1
DBD 26	TATA-binding proteins	1
DBD 27	Grainyhead domain factors	1
DBD 28	SAND domain factors	1
DBD 29	Psq-type HTH domain	1
DBD 30	CRC domain	1

Table 1: DNA-binding domains of the TFs within the JASPAR database.

A.1.2 ChIP-seq data preparation

To evaluate our method MASSIF, we downloaded 102 TF ChIP-seq datasets assayed in the human cell line K652 from ENCODE (The ENCODE Project Consortium, 2012). The ENCODE accession numbers are listed in Table 2. We used optimal IDR threshold peak calls in narrow bed format computed with the uniform ENCODE processing pipeline and version GRCh38 of the human reference genome. In the case that more than one peak file was available for a TF, we randomly chose one. Next, we used the BEDTools (Quinlan and Hall, 2010) getfasta functionality to extract the genomic sequences corresponding to the ChIP-seq peaks. In the header of the fasta file, we listed the genomic location of the sequences, extended by the *signalValue* provided in the bed files. If a tool that links motifs to TFs based on TF-associated sequences requires a biological information we used this *signalValue* provided by the ChIP-seq data. Sets of different sequence length are obtained from the middle of the peaks.

TF	accession number	TF	accession number
ZFX	ENCFF572OWY	JUN	ENCFF394CEC
MAFK	ENCFF002CXC	MAFF	ENCFF002CXB
MAFG	ENCFF393CCU	ZNF143	ENCFF002CYR
BHLHE40	ENCFF002CVQ	ZNF263	ENCFF002CYS
RELA	ENCFF931SVP	NFIC	ENCFF092TVM
SP1	ENCFF452LDK	CREB3	ENCFF606EUI
NR2C2	ENCFF789AXP	NR2C1	ENCFF664ZGR
GATA2	ENCFF002CMA	GATA1	ENCFF632NQI
RFX5	ENCFF002CXV	ARNT	ENCFF447FIO
RFX1	ENCFF010UHD	TCF7L2	ENCFF556FYF
FOXA1	ENCFF765NAN	RUNX1	ENCFF545WXN
TCF12	ENCFF912LXU	CTCF	ENCFF002CLS
HINFP	ENCFF558VAL	TCF7	ENCFF512IAI
ESRRA	ENCFF592GWM	BACH1	ENCFF365WQR
EGR1	ENCFF558JBX	MNT	ENCFF459DYU
ELK1	ENCFF002CWO	NFE2	ENCFF312XHI
ZBTB33	ENCFF002CMY	FOXJ2	ENCFF175IUD
CUX1	ENCFF556HMX	E2F7	ENCFF013EHI
PKNOX1	ENCFF062VBB	ETV1	ENCFF804KSB
IRF2	ENCFF886EVL	MGA	ENCFF525MPI
ETV6	ENCFF660VPM	TBP	ENCFF002CYK
IRF9	ENCFF863WCN	MEF2A	ENCFF002CMD
CREM	ENCFF021XJN	MEF2D	ENCFF257RYT
HES1	ENCFF010OOE	TFDP1	ENCFF600MGE
YY1	ENCFF002CMX	SMAD2	ENCFF186MFI
IKZF1	ENCFF994OQH	STAT2	ENCFF002CYG
STAT1	ENCFF002CYB	ZNF740	ENCFF301FKQ
NR2F6	ENCFF712AXK	NR2F1	ENCFF363IQN
USF2	ENCFF002CYP	USF1	ENCFF002CMV
FOSL1	ENCFF002CLY	ZBED1	ENCFF388TYU
TAL1	ENCFF002CYH	E2F6	ENCFF002CLU
IRF1	ENCFF002CWW	MAX	ENCFF002CXD
LEF1	ENCFF043YZF	MYC	ENCFF002CWD
NFYB	ENCFF002CXJ	SREBF1	ENCFF777MYW
GABPA	ENCFF002CLZ	NRF1	ENCFF782YFS
JUNB	ENCFF739XTO	JUND	ENCFF002CWZ
ELF1	ENCFF617ZLL	NFYA	ENCFF002CXI
MITF	ENCFF071NYD	MYBL2	ENCFF905KOD
SOX6	ENCFF431STY	ELF4	ENCFF539SXG
TEAD4	ENCFF002CMT	KLF13	ENCFF381GEK
CEBPB	ENCFF002CVV	CEBPG	ENCFF086CSF

Table 2 – continued from previous page

TF	accession number	TF	accession number
KLF16	ENCFF844HBQ	CTCFL	ENCFF002CLT
ETS2	ENCFF772HOY	ETS1	ENCFF002CLX
KLF1	ENCFF287GDT	ZBTB7A	ENCFF002CMZ
NR4A1	ENCFF859NPS	REST	ENCFF002CMF
PBX2	ENCFF269EMM	ARID3A	ENCFF002CVL
THAP1	ENCFF002CMU	E2F5	ENCFF468VJV
E2F4	ENCFF002CWM	E2F1	ENCFF445VTT
HMBOX1	ENCFF718DFX	CREB3L1	ENCFF566HGU
ATF7	ENCFF371SJR	ATF1	ENCFF030HWZ
ATF3	ENCFF002CLN	ATF2	ENCFF803FHN

Table 2: Accession numbers of the downloaded TF ChIP-seq dataset.

A.2 Further information to Chapter 4

In the following, for each kind of EpiRegio's REST API query a detailed instruction and an example is provided. Thereafter, the Python package *request* is used as an example how to access our database via HTTPS requests. Further, we outline the result of the literature search for the CRISPR -Cas analysis and the second CRISPR-Cas screen. Finally, we explain the details about our bioinformatics analysis within the circPLOD2 project.

A.2.1 Examples of the queries of the REST API

The queries of EpiRegio's REST API allow to retrieve the same information as the web interface provides. The general syntax of the queries is

```
https://epiregio.de/REST_API/<query>/<optionalParamters>/<input>/.
```

The result is given in JSON format.
GeneQuery: Given one or more ensembl IDs or gene symbols, the *GeneQuery* returns all REMs associated to the given genes. The following information is provided: geneID, geneSymbol, REMID, chr, start, end, regressionCoefficient, p-value, model score, version of the Epiregio database, number of REMs per CREM, CREM ID, and a list of the cell type or tissue score and the DNase1 signal of the cell type used in STITCHIT. In comparison to the web interface it is not possible to only select a subset of the used cell types or tissues, nor

to specify a cutoff for the celltype and tissue associated score/signal. In the following an example HTTPS request to retrieve the REMs for multiple genes is given:

```
https://epiregio.de/REST_API/GeneQuery/ENSG00000138232_ENSG00000274220/
```

RegionQuery: This query returns for one or several genomic regions REMs that overlap with at least one of the regions for a user-defined percentage. The genomic region is expected as *chr:start-end*, with $start < end$. The user-defined percentage of overlap can be specified optionally, the default is 100%. The same information as for the *GeneQuery* is given as output. The following HTTPS request, returns REMs lying completely within the given regions:

```
https://epiregio.de/REST_API/RegionQuery/chr16:75423800-75424410_chr2:1369428-1369900/
```

The second example provides the resulting REMs that overlap at least 50% with the given region:

```
https://epiregio.de/REST_API/RegionQuery/50/chr16:75423948-75424405/.
```

REMQuery: Given one or several REM IDs, one can look up all information stored about the given REMs, like the associated target genes, the genomic position or the model score. The input is expected to be a valid REM ID. The output format is the same as for the *GeneQuery* and the *RegionQuery*. In the following an example with three REM IDs is shown:

```
https://epiregio.de/REST_API/REMQuery/REM0000011_REM0000006_REM00000111/
```

CREMQuery: This query lists all REMs contained within one or several given CREM IDs. The given information per REM are the same as for the GeneQuery. An exemplary HTTPS request is given below:

```
https://epiregio.de/REST_API/CREMQuery/CREM0000132_CREM0000018/
```

GeneInfo: Given one or several Ensembl IDs, general information about the genes *e.g.* the genomic position, the gene symbol or the strand is returned. For the genes *ENSG00000223977* and *ENSG00000022397* the information can be retrieved as following:

```
https://epiregio.de/REST_API/GeneInfo/ENSG00000223977_ENSG00000223979/
```

```
import requests

regions = ['ENSG00000275675', 'ENSG00000123201', '
    ENSG00000138231'] # genes of interest
results = []  # Listi, where the output is stored

# loop over all genes
for g in genes:
        url = 'https://epiregio.de/REST_API/GeneQuery/' + g
    + '/'
        api_call = requests.get(url)
        if api_call.status_code != 200:  # In case the page
    does not work properly.
                print("Page Error")
        for hit in api_call.json():
                results.append([hit['REMID'], hit['chr'],
    hit['start'], hit['end'], hit['cellTypeScore']['heart'
    ]])

print(results)
```

Code 1: Accessing EpiRegio's REST API via Pythons package Requests. Example showing how to retrieve for a given gene the associated REMs

A.2.2 Programmatic access to the REST API via Python

To regularly extract data from our website or to include EpiRegio's queries within a script, it is possible to call our REST API outside of a browser. To do so, a program is needed that allows to build HTTPS requests, for instance the Python package Requests. In Code 1 an example is shown, where the REMs associated to the genes *ENSG00000275675*, *ENSG00000123201* and *ENSG00000138231* are gathered and the REM ID, the genomic position and the cell type or tissue score for heart are given as output.

A.2.3 CRISPR-Cas analysis: Genes associated to doxorubicin resistance, chemoresistance against other drugs or cancer growth

For the genes, for which we found literature evidence to be associated to chemoresistance against doxorubicin or other cancer drugs as well as to cancer growth are outlined below (see also Table 4.1 in Chapter 4).

***MMP9* and *NCOA5*:** Myocytes treated with doxorubicin show an increasing ROS formation leading to an increase of Myocardial Metallo Proteinases

(MMPs) expression, namely MMP2 and *MMP9* (Spallarossa *et al.*, 2006). These findings are in agreement with our prediction that a REM modified by a gRNA on chromosome 20 is a regulatory site of *MMP9*. In addition, it was shown that *NCOA5* was upregulated in colorectal cancer leading to a higher expression of *MMP9* and Cyclin D1, whereas the expression of p27 was lower through PI3K/AKT pathway. This enhances cell proliferation, migration and invasion (Sun *et al.*, 2017).

***NT5C2*:** Mutations in the gene *NT5C2* are known to lead to treatment resistance against thiorubin in acute lymphocytic leukemias (Dieck and Ferrando, 2019).

***ANXA7*:** Liu *et al.* (2020) observed that *ANXA7* expression is increased in multiple myeloma cells. They hypothesize that *ANXA7* promotes cell cycle, proliferation and cell adhesion-mediated drug resistance via the upregulation of CDC5L. Another study analyzing diffuse large B-cell lymphoma with a weighted gene co-expression network identified *ANXA7* as associated with overall survival of patients treated with chemotherapy including doxorubicin among other drugs (Gao *et al.*, 2020).

***TMEM219*, *GDPD3*, and *DOC2A*:** A gRNA overlapping with active REMs on chromosome 16 is of particular interest since we discovered literature evidence for 3 out of 4 associated genes. It has been shown that *TMEM219*, a IGFBP-3 receptor, is associated to promote antitumoric effects of IGFBP-3 (Cai *et al.*, 2020). The expression of *GDPD3* as well as the protein abundance of *GDPD3* is linked to a positive response to neoadjuvant chemotherapy in urothelial carcinomas (Baras *et al.*, 2015). A genome-wide screen for induced genes after doxorubicin treatment in cancer cell lines (Indermaur *et al.*, 2010) identified that among other genes *GDPD3* and *DOC2A* showed differential expression.

***TMEM199*:** In cisplatin resistent oesophageal adenocarcinoma cells the gene expression of *TMEM199* was upregulated, pinpointing to its potential role in chemoresistance (Buckley *et al.*, 2019).

***SLC25A10*:** Potential cancer therapy targets are proteins that belong to the family of mitochondrial carriers, to which the product of the gene *SLC25A10* belongs to (Rochette *et al.*, 2020). It was shown, that a high protein level of *SLC25A10* is linked to high metastasis potential as well as a lower relapse-free survival in osteosarcomas (Wang *et al.*, 2020).

***PFKP*:** Moon and others found that PFKP, which is regulated by KLF4 is involved in cell proliferation in breast cancer cells (Moon *et al.*, 2011). Further, in lung cancer it was observed that high expression levels of *PFKP* is correlated with tumor aggressiveness and growth (Shen *et al.*, 2020).

MGTS2: Endoplasmic reticulum stress and doxorubicin or 5-fluorouracil treatment induce the expression of MGTS2, and by that LTC4, which triggered DNA damage and cell death (Dvash *et al.*, 2015). Dvash *et al.* demonstrated

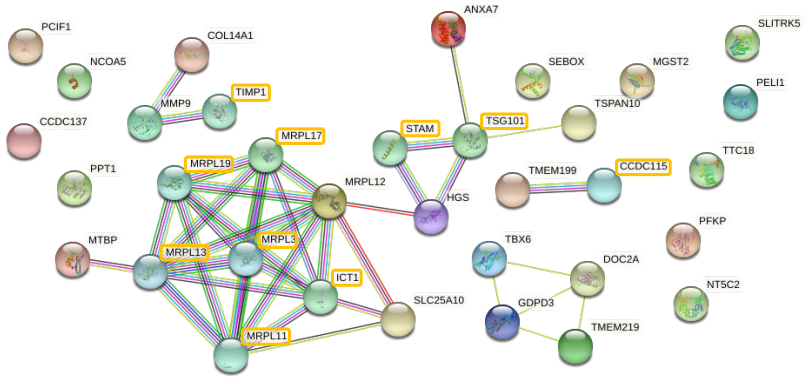


Figure 1: Functional analysis of candidate genes. Protein-protein association network of the genes associated to active, genomically modified REMs derived using the STRING database (Szklarczyk *et al.*, 2020). The network includes all genes from Table 4.1 and 10 additional interactors/proteins (marked with orange box) highly connected to the genes provided as input.

that MGTS2-deficient mice are resistant against 5-fluorouracil treatment. In addition, they hypothesized that in cell lines which do not express *MGTS2* commonly used cancer drugs like doxorubicin or 5-fluorouracil can not activate the MGST2-LTC4 pathway to trigger DNA-damage and cell death.

A.2.4 CRISPR analysis: Protein-protein association network

In Figure 1 the resulting protein-protein interaction network for the 35 associated genes from Table 4.1 in Chapter 4 is shown.

A.2.5 CRISPR-Cas analysis: Literature evidence for TFs enriched within the REMs of candidate target genes

In the following the association between the enriched TFs and chemoresistance against doxorubicin or other cancer durgs as well as cancer growth are explained in more detail (see also Table 4.2 in Chapter 4).

***NR2C2* and *NR2F2*:** For *NR2C2* (or testicular nuclear receptor TR4) it has been shown that it can suppress cancer invasion by the prevention of infiltration of macrophages (Hu *et al.*, 2019). Further, in the same publication it was demonstrated that the sensitivity of docetaxel chemotherapy in prostate cancer can be enhanced by utilizing antagonists against *NR2C2*. Further, the TF *NR2F2* (or Chicken ovalbumin upstream promoter transcription factor II,

COUP-TF2), which is expressed in different types of colorectal carcinoma, is known to promote the resistance to doxorubicin (Wang *et al.*, 2016). We found several reviews that point to the important role of nuclear receptors during the development of cancer like breast or prostate cancer, and the potential to target them with cancer drugs (Zhao *et al.*, 2019; Riggins *et al.*, 2010; Wu *et al.*, 2016). Within our analysis we identified not only *NR2F2* and *NR2C2* factors, but also RARA, NR3C1, NR2C1 and THRB as nuclear receptors.

***SREBF1* and *SREBF2*:** Recently, G protein-coupled receptor 120 (GPR120) was identified to be a chemoresistance promoting factor in breast cancer (Wang *et al.*, 2019). GPR120 mediates fatty acid synthesis and within this process the authors identified several factors, like *SREBF1* and *SREBF2*. As a consequence of the increased lipid synthesis, the cell membrane has a different lipid composition, which may prevent the drug from being absorbed.

***RUNX3*:** In human lung adenocarcinoma it has been shown that the down regulation of *RUNX3* is linked to docetaxel resistance when Akt1-mediated signaling is activated (Zheng *et al.*, 2013). When *RUNX3* is overexpressed, the AKT expression is decreased, and a higher sensitivity for cancer drug treatment is observed.

***ATF3*:** The TF *ATF3* is well known to play a role in cellular stress response and is therefore enriched in cells exposed to stress signals (Hai *et al.*, 1999). Further, it has been reported that under doxorubicin treatment *ATF3* is involved in cell death and cell cycle progression (Nobori *et al.*, 2002; Park *et al.*, 2012). However, an open question is whether *ATF3* works as a negative or positive regulator.

***ATF2*, *ATF7* and *AP-1*:** The TFs *ATF2* and *ATF7* are homologous members in the activator protein 1 (AP-1) family. In mouse it has been shown that MYC-expressing lymphoma cells are resistant against doxorubicin induced apoptosis when the TFs *ATF2* and *ATF7* is lost (Walczynski *et al.*, 2013). In addition, the authors hypothesize that *ATF2*/*ATF7* play an important role in suppressing oncogenic transformation. Another study found that upregulating *AP-1* leads to resistance to cancer drugs (Wang *et al.*, 2018b)

A.2.6 CRISPR-Cas analysis: Evaluation of two additional screens

In this section, we outline the details of the results of the additional screens (both published in (Klann *et al.*, 2017)) we tested our analysis protocol on.

Analysis of the β-globin locus: Klann *et al.* (2017) studied the β-globin locus in a K562 myelogenous leukemia cell line. This locus incorporates 5 globin genes of which 3 (HBE1, HBG1 and HBG2) were highly expressed in the studied cell line. Further, 5 DNase1 hypersensitive sites (DHSs) were identified upstream of the β-globin locus, which fall within the region that is known to control the expression of the β globin genes in the studied locus (Levings

and Bungert, 2002). The authors identified 7 gRNAs that either overlap a DHS or the promoter of HBG1 or HBG2 and thereby significantly change the gene expression of HBE1. We applied our analysis pipeline to determine potential target genes, and identified several REMs associated to HBE1. Further, among the associated genes we observed the genes HBG1, HBG2, HBD, and HBB, which are located within the β-globin locus, and genes from the olfactory receptor gene 51 subfamily B, namely OR51B2, OR51B4, OR51B5 and OR51B6.

Analysis of gRNA target sites effecting HER2 expression Next, we applied our analysis pipeline to 31 gRNA target sites significantly increase or decrease the expression of HER2 (synonym ERRB2) identified with a screen done in the HER2-positive cancer cell line SKBR3. Among the 11 target genes we detected HER2. Further, the gene GRB7 was identified, which is co-expressed with HER2 in some breast cancer cells (Nadler *et al.*, 2010).

A.2.7 CircPLOD2 analysis: Identification of REMs linked to DEGs and TF motif enrichment analysis

Using EpiRegio's Gene Query we extracted 68.087 REMs for the top 1000 DEGs after circPLOD2 depletion. The DEGs are sorted by the adjusted p-value (Benjamini-Hochberg correction) collected from DESeq2 result. The DNase1-seq data to check for active REMs and to sort the REMs according to the DNase1 signal in human brain pericytes was taken from the ENCODE portal (The ENCODE Project Consortium, 2012; Luo *et al.*, 2019) with the identifier *ENCFF353QSZ*. For the 16.901 active REMs, the DNA sequence was obtained using bedtools getfasta functionality (version 2.27.1) (Quinlan and Hall, 2010) for the human genome version hg38. We downloaded 633 human TF binding motifs from the JASPAR database (version 2020) (Fornes *et al.*, 2019) and excluded those not or weakly expressed in our RNA-seq data (TPM <= 1), resulting in 292 TFs. For the TF motif enrichment analysis we used *PASTAA*, which identified 95 TF significantly enriched (adjusted p-value less or equal 0.01, Benjamini-Hochberg correction). Next, we determined the TF binding sites of the enriched TFs within the REMs using FIMO (version meme-5.2.0) (Grant *et al.*, 2011). Then, we counted the number of predicted binding site for each TF and for each REM and visualized the counts as a heatmap utilizing R's complexHeatmap package. A slightly modified workflow of this analysis can be found on our EpiRegio GitHub repository: https://github.com/Team Regio/ApplicationScenarioExamples/tree/master/PASTA_Fimo_analysis.

A.2.8 CircPLOD2 analysis: Computation of chromatin accessibility profiles

Using deeptools (Ramírez *et al.*, 2016), we visualized the DNase1 signal of REMs with at least one predicted KLF4 binding site (3,025 of the 68,086 REMs). We extracted the positions of the REMs, and extended them to 4 kb long regions with the REM centered in the middle. Additionally, we required the DNase1 signal from the human brain pericytes used for the TF motif enrichment analysis. The following two commands were used:

```
computeMatrix reference-point -S <DNaseSignal.bigWig> -R <
   REMs> -a 2000 -b 2000 -o <output.tab.gz> - -
   referencePoint center
plotHeatmap -m <output.tab.gz> -o <heatmap.pdf> --
   regionsLabel REMs.with.KLF4.binding.site --
   legendLocation none --samplesLabel DNase-signal -x '
   distance (bp)'
```

A.3 Further information to Chapter 5

In this section, we provide the commands used to run our statistical approach and atSNP.

A.3.1 Utilized commands to apply our approach and *atSNP*

For a better reproducibility of our results, we provide all executed commands and the input files in a ZENODO repository (https://doi.org/10.5281/zenodo.7588272, for details see Section A.5.2). For all analyses we used the genome version hg38, besides for the SNP-SELEX data where we used hg19. The dataset specific input files are indicated with <>. The SNP and motifs data for the different dataset for our approach and *atSNP* are provided in their required format, the content is the same.

Commands to apply our statistical approach

In the following the used commands to run our approach are listed. All scripts and more details on how to run them can be found in our GitHub repository: https://github.com/SchulzLab/SNEEP.

As a first step, we need to estimate the scale parameter b for the used PWMs using

```
bash estimateScalePerMotif.sh 200000 <motifs> <outputDir> <
    motifNames> 1.9
```

As *motifs* we used the file *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt*, which can be obtained from our ZENODO data repository. *motifNames* lists the names of all TF motif for which we wish to compute b. The result is a file providing for each considered TF model the estimated scale parameter b, called *estimatedScalesPerMotif_1.9.txt* in the following.

To compute the results for the ASB events, the SNP-SELEX data and the runtime analyses we used the following command:

```
time ./src/differentialBindingAffinity_multipleSNPs -o <
    outputDir> -n <numberCores> -p 1.0 -c 1.0   -s
    estimatedScalesPerMotif_1.9.txt  -j 0 <motifFile> <input
    -snps>  <path-to-genome-file>.
```

The file *motifFile* can be found in our ZENODO data repository as well as the *input-snps*:

- ASB events: motifs_ASB.txt, snps_ASB.txt
- SNP-SELEX data: motifs_SNP_SELEX.txt, snps_SNP_SELEX.txt
- randomly sampled data: motifs: *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt* and the randomly sampled SNPs can be found in the files sampledSNPsi.txt with $i \in \{100, 500, 1,000, 10,000, 20,000, 40,000\}$.

A.3.2 Commands used to run *atSNP*

To run *atSNP* (version 1.14.0) the following commands where executed in R:

```
pwms <- LoadMotifLibrary(<motifFile>, tag = "MOTIF",
    skiprows = 2, skipcols = 0, transpose = FALSE, field =
    2, sep = " ", pseudocount = 0)
snps <- LoadSNPData(filename = <snpFile>, genome.lib = <
    genomeVersion>)
scores <- ComputeMotifScore(pwms, snps, ncores = <
    numberCores>)
diffBind <- ComputePValues(motif.lib = pwms, snp.info = snps
    ,  motif.scores = scores$motif.scores, ncores = <
    numberCores>)
```

The used motif input files and the SNP files are uploaded to our ZENODO data repository:

- ASB events: motifs_ASB_atSNP.txt, snps_ASB_atSNP.txt,

- SNP-SELEX data: motifs_SNP_SELEX_atSNP.txt, snps_SNP_SELEX_atSNP.txt
- randomly sampled data: motifs: motifs_JASPAR_HOCOMOCO_Kellis_1.9.meme and the randomly sampled SNPs can be found in the files sampledSNPsi_atSNP.txt with $i \in \{100, 500, 1.000, 10,000, 20,000, 40,000\}$.

As *genomeVersion* we used "BSgenome.Hsapiens.UCSC.hg38" for the ASB events and "BSgenome.Hsapiens.UCSC.hg19" used for the SNP-SELEX data.

A.4 Further information to Chapter 6

A.4.1 A summary report for the GWAS atherosclerosis

The atherosclerosis GWAS was download from the GWAS Catalog with the ID EFO_0003914, including the child traits *carotid atherosclerosis* (EFO_0003914) and *peripheral arterial disease* (EFO_0009783). We excluded all indels, duplicate SNPs and all SNPs from GWAS not based on a European cohort. For the remaining 255 out of 261 SNPs, we determined the SNPs in linkage disequilibrium (LD) applying SNiPA (Arnold *et al.*, 2014) using as population the European cohort and an LD threshold of 0.8. We gathered 4,326 unique proxy SNPs. We computed the D_{max} p-value for each collected SNP associated to atherosclerosis. As input motifs we used 817 non-redundant human motifs from the JASPAR (version 2022), HOCOMOCO and Kellis ENCODE motif database. We ran our Sneep workflow with following command:

```
time ./src/differentialBindingAffinity_multipleSNPs -o <
    outputDir>  -n 10 -p 0.5 -c 0.001  -r REM.txt -s
    estimatedScalesPerMotif_1.9.txt  -g ensemblID_GeneName.
    txt  -j 1000 -l 10 -k dbSNPs_sorted.txt
    combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar2022
    .txt  <input-snps.txt>  <path-to-genome-file>
```

The files *interactionsREMs.txt*, *ensemblID_GeneName.txt* and *dbSNPs_sorted.txt* can be found in our github repository and in our ZENODO data repository, the SNPs of the GWAS atheroclersis are stored in the file snps_atherosclerosis.txt and the motifs are provided in the file *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt*. If the D_{max} p-value is ≤ 0.001, we assumed that the binding site of a TF was significantly affected by the considered SNP. We linked the rSNPs to genes using the REM-gene interactions from EpiRegio (Baumgarten *et al.*, 2020; Schmidt *et al.*, 2021). Parts of the resulting summary report of our Sneep workflow are visualized in Figure 2.

A.4.2 eQTL application: Data collection and used commands

We gathered the eQTLs for *EBV-transformed lymphocytes* and *cultured fibroblasts* from the GTEx portal (GTEx Consortium, 2017) (version 8; dbGaP Accession phs000424.v8.p2) on December 21, 2022. The fine-mapped eQTLs were extracted from the file *GTEx_v8_finemapping_CaVEMaN.txt.gz* (Brown *et al.*, 2017) resulting in 14,722 eQTLs for lymphocytes and 45,917 eQTLs for fibroblasts. Further, we downloaded the genes transcript per million (TPM) values of lymphocytes and fibroblasts from the GTEx portal (*gene_tpm_2017-06-05_v8_cells_cultured_fibroblasts.gct.gz* and *gene_tpm_2017-06-05_v8_cells_ebv-transformed_lymphocytes.gct.gz*).
We ran SNEEP including the TF enrichment statistic separately for each cell type using the following command:

```
time ./src/differentialBindingAffinity_multipleSNPs -o <
    outputDir>  -n 16 -p 0.5 -c 0.01 -s
    estimatedScalesPerMotif_1.9.txt -j 1000  -l 10 -k
    dbSNPs_sorted.txt
    combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar2022
    .txt <input-snps>  <path-to-genome-file>
```

The *input-snps* are either the eQTLs associated to lymphocytes or fibroblasts, the file *dbSNPs_sorted.txt* is part of our github repository and the file *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt* can be found in the SNEEP ZENODO data repository.
We list the TFs which are more often affected by the eQTLs than expected for lymphocytes in Table 3 and for fibroblasts in Table 4.

geneID	gene name	odds-ratio
ENSG00000163884	KLF15	4.198182636
ENSG00000198081	ZBTB14	3.701877107
ENSG00000172273	HINFP	3.557608721
ENSG00000168267	PTF1A	3.450369597
ENSG00000069812	HES2	3.311657814
ENSG00000106459	NRF1	3.091872126
ENSG00000217236	SP9	2.954418515
ENSG00000175832	ETV4	2.927260312
ENSG00000164683	HEY1	2.883928101
ENSG00000169057	MECP2	2.86899042
ENSG00000100644	HIF1A	2.82095553
ENSG00000183733	FIGLA	2.811376843
ENSG00000139800	ZIC5	2.746991856
ENSG00000111145	ELK3	2.727669997

ENSG00000163848	ZNF148	2.724808458
ENSG00000109787	KLF3	2.662947757
ENSG00000134046	MBD2	2.659966678
ENSG00000120738	EGR1	2.564840185
ENSG00000007968	E2F2	2.552071613
ENSG00000101412	E2F1	2.542092192
ENSG00000124092	CTCFL	2.541007285
ENSG00000177485	ZBTB33	2.517629911
ENSG00000172845	SP3	2.499383043
ENSG00000158711	ELK4	2.496004584
ENSG00000173404	INSM1	2.471286487
ENSG00000135547	HEY2	2.462675967
ENSG00000100105	PATZ1	2.460509115
ENSG00000197714	ZNF460	2.457716009
ENSG00000155090	KLF10	2.446171012
ENSG00000009950	MLXIPL	2.410562045
ENSG00000104903	LYL1	2.405839145
ENSG00000172059	KLF11	2.403109656
ENSG00000184635	ZNF93	2.401127541
ENSG00000087510	TFAP2C(MA0815.1)	2.316142954
ENSG00000067082	KLF6	2.305237661
ENSG00000266265	KLF14	2.289781338
ENSG00000134954	ETS1	2.275715041
ENSG00000105722	ERF	2.272464792
ENSG00000181315	ZNF322	2.248071038
ENSG00000184937	WT1	2.233055339
ENSG00000174738	NR1D2	2.221471711
ENSG00000126368	NR1D1	2.219677798
ENSG00000087510	TFAP2C(MA0524.2)	2.218377391
ENSG00000127528	KLF2	2.211052685
ENSG00000139352	ASCL1(MA1100.2)	2.192946519
ENSG00000116819	TFAP2E	2.189332628
ENSG00000008196	TFAP2B(MA0813.1)	2.183194328
ENSG00000198146	ZNF770	2.145992427
ENSG00000188868	ZNF563	2.142626064
ENSG00000205250	E2F4	2.142299562
ENSG00000118922	KLF12	2.127033879
ENSG00000008196	TFAP2B(MA0811.1)	2.114145685
ENSG00000105866	SP4	2.108886473
ENSG00000167182	SP2	2.100411031
ENSG00000101190	TCFL5	2.099437045
ENSG00000092607	TBX15	2.091189665

ENSG00000137203	TFAP2A(MA0872.1)	2.073639795
ENSG00000164651	SP8	2.069864364
ENSG00000107249	GLIS3	2.065533602
ENSG00000156925	ZIC3	2.065533566
ENSG00000114315	HES1	2.05645807
ENSG00000126767	ELK1	2.042948778
ENSG00000118263	KLF7	2.036872895
ENSG00000136826	KLF4	2.033989175
ENSG00000008196	TFAP2B(MA0812.1)	2.016281574
ENSG00000101216	GMEB2	2.003061144

Table 3: Enriched TFs observed in lymphocytes. The TFs more often affected than expected on random data (odds-ratio > 2) are listed.

geneID	gene name	odds-ratio
ENSG00000163884	KLF15	4.133870339
ENSG00000106459	NRF1	3.844084087
ENSG00000198081	ZBTB14	3.797638419
ENSG00000178187	ZNF454	3.163969949
ENSG00000100105	PATZ1	3.034508041
ENSG00000177485	ZBTB33	2.957630324
ENSG00000217236	SP9	2.950333517
ENSG00000204644	ZFP57	2.851643008
ENSG00000169057	MECP2	2.82148639
ENSG00000124092	CTCFL	2.783070458
ENSG00000164683	HEY1	2.754913691
ENSG00000167182	SP2	2.691105222
ENSG00000205250	E2F4	2.630842718
ENSG00000185669	SNAI3	2.625177341
ENSG00000007968	E2F2	2.605868172
ENSG00000183733	FIGLA	2.594100325
ENSG00000139800	ZIC5	2.571281633
ENSG00000067082	KLF6	2.514779852
ENSG00000103495	MAZ	2.51440285
ENSG00000135547	HEY2	2.512672434
ENSG00000184635	ZNF93	2.511454592
ENSG00000179111	HES7	2.501558836
ENSG00000114315	HES1	2.477429462
ENSG00000134046	MBD2	2.435649223
ENSG00000179388	EGR3	2.41086929

ENSG00000173480	ZNF417	2.377937061
ENSG00000118263	KLF7	2.359097513
ENSG00000133794	ARNTL	2.356421326
ENSG00000137203	TFAP2A(MA0872.1)	2.342720217
ENSG00000124216	SNAI1	2.289609986
ENSG00000204335	SP5	2.281799418
ENSG00000102974	CTCF(MA0139.1)	2.266367443
ENSG00000102554	KLF5	2.250960463
ENSG00000136451	VEZF1	2.250263322
ENSG00000172059	KLF11	2.243126545
ENSG00000127528	KLF2	2.237589531
ENSG00000108788	MLX	2.227134188
ENSG00000184937	WT1	2.225847249
ENSG00000266265	KLF14	2.221953878
ENSG00000101216	GMEB2	2.21485651
ENSG00000181894	ZNF329	2.20037388
ENSG00000177551	NHLH2	2.181538801
ENSG00000105866	SP4	2.169468655
ENSG00000160685	ZBTB7B	2.161818174
ENSG00000005889	ZFX	2.15990653
ENSG00000109787	KLF3	2.148801004
ENSG00000137203	TFAP2A(MA0810.1)	2.145871091
ENSG00000068323	TFE3	2.131141099
ENSG00000162702	ZNF281	2.124542884
ENSG00000172845	SP3	2.123593439
ENSG00000008196	TFAP2B(MA0812.1)	2.076304533
ENSG00000008196	TFAP2B(MA0813.1)	2.07242496
ENSG00000198146	ZNF770	2.069549246
ENSG00000111049	MYF5	2.069301566
ENSG00000101190	TCFL5	2.052977374
ENSG00000164651	SP8	2.024413763
ENSG00000156273	BACH1	2.021032426

Table 4: Enriched TFs observed in fibroblasts. The TFs more often affected than expected on random data (odds-ratio > 2) are listed.

A.4.3 CVD associated ncRNAs: Collection of GWAS SNPs of cardiovascular diseases

We have collected the significant SNPs from the NHGRI-EBI GWAS Catalog (Buniello *et al.*, 2018) for the following search terms including all the

available child traits: Coronary artery disease (EFO_0000378), Aortic stenosis (EFO_0000266), Cardiac arrhythmia (EFO_0004269), Cardiomyopathy (EFO_0000407), Myocardial infarction (EFO_0000612) and Myocardial ischemia (EFO_0005672). All GWAS were downloaded on 10/26/2020. For each set of GWAS SNPs, we have obtained correlated SNPs that are in linkage disequilibrium (LD) with any of the original SNPs. We used the LDProxy Tool (Machiela and Chanock, 2015) and extracted the proxy SNPs via their API functionality. Proxy SNPs with an $R^2 \leq 0.75$ and within a window of +/- 500.000 bp centered around the original SNP were added to the GWAS SNPs. The combined set of proxy and lead SNPs was used as input set to the SNEEP workflow.

A.4.4 CVD associated ncRNAs: Detection of regulatory SNPs

To detect regulatory SNPs and associated target genes, we applied the SNEEP workflow with the optional functionality to integrate REM-gene interactions separately for each of the 6 cardiac GWAS. In the following the used commands to run the SNEEP workflow are explained in detail.
In a first step, we estimated the scale parameter b separately for each motif using the script *estimateScalePerMotif.sh*. Therefore, we sampled 200,000 SNPs randomly from the dbSNP database (Sherry, 2001) and removed flanking bases of the PWMs with an entropy higher than 1.9, resulting in the following command:

```
bash estimateScalePerMotif.sh 200000  <pathToMotifs> <
    outputDir> <motifNames> 1.9
```

To run SNEEP, 632 human TF PWM motifs in TRANSFAC format were gathered from the JASPAR database (version 2020) (Fornes *et al.*, 2019) and over 2.4 million regulatory elements linked to their putative target genes were downloaded from the EPIREGIO database (https://doi.org/10.5281/zenodo.3758494, file: REMAnnotationModelScore_1.csv.gz). We applied the SNEEP workflow per GWAS with a D_{max} p-value cutoff of 0.001:

```
./differentialBindingAffinity_multipleSNPs -o <
    OutputDirectory>  -n 10 -p 0.5 -c 0.001 -r  <REMs-gene-
    interactions>  -g  <EnsemblMapping>  -j 100  -l 123 -i <
    pathToSneepSoftware>  -s <estimatedScales>  <Motifs>  <
    InputSnps>  <hg38>
```

A.4.5 CVD associated ncRNAs: Identification of disease associated genes using rSNPs

As part of the analysis of SNEEP, all rSNPs that overlap regulatory elements from EPIREGIO webserver constitute a candidate disease gene. From the SNEEP output file protein-coding and non-coding genes were extracted that have overlapping rSNPs in their regulatory elements. Non-coding genes were understood to be all genes not labeled with the biotype protein coding or TEC (primary assembly annotation, version 39, downloaded from GENCODE). To label which protein-coding genes are already associated to the studied diseases (Figure 6.4B, Chapter 6, black dots), we used the disease2gene functionality of the R package of DisGeNET:

```
disease2gene(disease = <studiedDisease>, database = "ALL"
, score = c( 0,1))
```

The studiedDisease parameter needs to be provided as UMLS CUI identifiers. We used C1956346 (CAD), C0003811 (Cardiac arrhythmia), C1449563 (Cardiomyopathy, Familial Idiopathic), C0151744 (Myocardial ischemia), C0027051 (Myocardial infarction) and C0340375 (Subaortic stenosis).

A.4.6 CVD associated ncRNAs: Disease enrichment analysis based on DisGeNET

We performed a disease enrichment analysis for the protein-coding genes, as well as for the non-coding genes. The protein coding genes could be directly use as input to the analysis. For the non-coding genes we conducted a co-expression analysis using the gene expression profiles of 9,662 GTEx RNA-seq samples. The joint set of the protein-coding genes that are co-expressed to any of the non-coding genes found for the same disease via the GWAS analysis, were used as input to the disease enrichment analysis.
The disease enrichment analysis was done separately for each set of genes either directly associated via rSNPs or in case of the non-coding genes we utilized the resulting co-expressed protein-coding gene sets. We used the function *disease_enrichment* from the R package of DisGeNET to perform the enrichment analysis:

```
disease_enrichment( entities = <coExpressedPCG>,
vocabulary = "HGNC", database = "ALL" )
```

The results of the analysis for the protein-coding genes is visualized in Figure 3. For the GWAS Myocardial ischemia and Aortic stenosis the disease enrichment analysis was not possible, because only 5 and 12 protein coding genes were associated, respectively. The result of the analysis for the co-expressed protein coding genes of the non-coding genes are visualized in Figure 6.5C in Chapter 6,

where cardiovascular disease associated phenotypes were selected and visualized as a dot plot using ggplot2. For the GWAS Aortic stenosis, Myocardial infarction and Myocardial ischemia we did not apply the disease enrichment analysis because of the low number of associated non-coding genes (less than 30).

A.5 ZENODO repositories

In the following the ZEBNODO repositories of EpiRegio and Sneep are outlined in more detail.

A.5.1 Details of the data stored in our EpiRegio webserver

In the ZENODO repository of our EpiRegio webserver available at https://doi.org/10.5281/zenodo.3750929, we provide the tables holding the data the webserver relies on. The content is 10 tables:

- GenomeAnnotation: contains information about genomeVersion, annotationVersion and databaseName (GenomeAnnotation_1.csv.gz)
- GeneAnnotation: Information of the genes (chr, start, end, geneID, geneSymbol, alternativeGeneID, isTF, strand and annotationVersion) (GeneAnnotation_1.csv.gz)
- GeneExpression of Blueprint and Roadmap: Per consortium one table containing information about geneID, sampleID, expressionLog2TPM and species (GeneExpression_Blueprint_1.csv.gz and GeneExpression-Roadmap_1.csv.gz)
- CellTypeInfo: Information of the used cell and tissue types (cellTypeID, cellTypeName and cellOntologyTerm) (CellTypeInfo.csv.gz)
- sampleInfo of Roadmap and Blueprint: Per consortium one table containing information about sampleID, originalSampleID, cellTypeID, origin and dataType (sampleInfo_Blueprint_1.csv.gz and sampleInfo_ Roadmap_1.csv.gz)
- REMAnnotation: contains all predicted REMs using STITCHIT (chr, start, end, geneID, REMID, regressionCoefficient, pValue, normModelScore, meanDNase1Signal, sdDNase1Signal, consortium and version) (REMAnnotationModelScore_1.csv.gz)

- REMActivity: This table contains for each REM the DNase1 signal and the standardised DNase1 signal per cell or tissue type (REMID, sampleID, dnase1Log2, standDnase1Log2 and version) (REMActivity_1.csv.gz)

- clusterREMs: contains all CREMs (REMID, CREMID, chr, start, end, REMsPerCREM and version) (clusterREMs_1.csv.gz)

A.5.2 Details about the utilized data within our statistical approach to identify rSNPs and Sneep

We provide a ZENODO data repository available at https://doi.org/10.5281/zenodo.7588272 for our preprint *A statistical approach to identify regulatory DNA variations*. The data is used within Chapter 5 and 6. The content is the following:

- ASB events and the corresponding PWMs used to evaluate our approach (for our approach and *atSNP* in their required format, for details see Section A.3.1).

- Result of our approach and *atSNP* for the ASB events (data for Figure 5.3A) (ASB_result_sneep_atSNP.txt)

- SNP-SELEX data used to evaluate our approach (for our approach and *atSNP* in the required format, for details see Section A.3.1).

- Result of our approach and *atSNP* for the SNP-SELEX (data for Figure 5.3B) (SNP_SELEX_result_sneep_atSNP.txt)

- Files containing the enriched TFs (odds-ratio > 2) for lymphocytes (enriched_TF_lymphocytes.txt) and fibroblasts (enriched_TF_fiborblasts .txt) (visualized in Chapter 5 in Figure 6.2)

- Input SNPs of the GWAS atherosclerosis (snps_atherosclerosis.txt) and the result of our approach (atherosclerosis_result.txt) (see Chapter 6, Section 6.1.6).

- TF input motifs in TRANSFAC format for our method and in MEME format for atSNP (see Section A.3.1)

A

TF	observed TF hits	average number TF hits sampled rounds	odds-ratio
PTF1A	17	7.577	2.248957
IRF5	16	6.724	2.385094
SOX17	14	6.927	2.024675
HOXA7	11	4.804	2.293327

B

TF	gain	loss	p-value	function
ZNF770	9	37	0.0000	loss
MSGN1	0	8	0.0078	loss
PRDM9	10	25	0.0167	loss
ZFP82	3	13	0.0213	loss
ZBTB7C	9	1	0.0215	gain

C

gene name	ensembl ID	number different interactions	number different SNPs	SNPs	TFs
RN7SL622P	ENSG00000265052	3	7	rs145520076	SREBF1(MA0829.2), ESRRG
				rs183763906	PRDM9
				rs189628481	ZNF770
				rs192988621	NFKB1
				rs200026262	ZNF554
				rs7214119	SP8, SALL4, TFAP2A(MA0810.1), ZNF281, WT1, ZNF467
				rs8075764	TBX18

Figure 2: Parts of the tables of the summary report. An atherosclerosis GWAS was analyzed using our SNEEP workflow including the TF enrichment statistic. Further, the rSNPs were linked to target genes using EPIREGIO. (**A**) - (**C**) hold the first few rows of the tables provided within the summary report. (**A**) The enriched TFs for the given SNP set are shown. For each TF we provide the number of observed TF hits on the input and on the sampled data as well as the resulting odds-ratio. (**B**) The TFs which are significantly associated to a gain or loss of function are visualized. (**C**) The target genes of the rSNPs are deciphered.

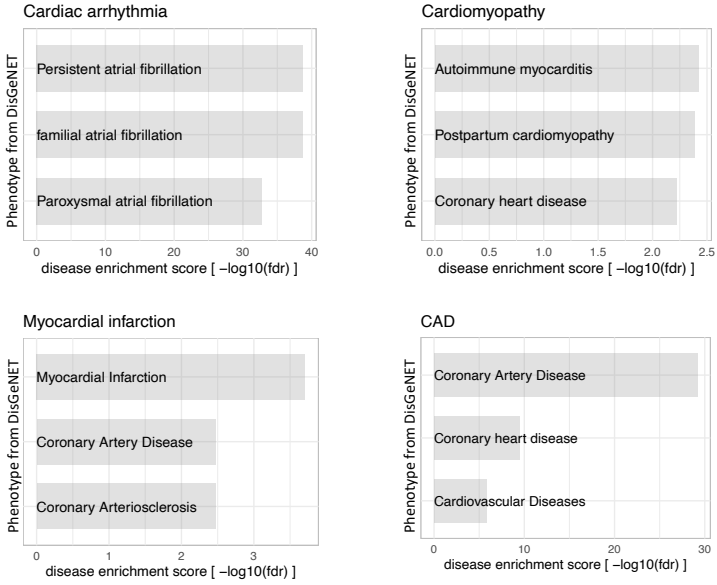


Figure 3: Disease enrichment analysis for protein-coding genes associated to cardiovascular diseases. Barplots representing per GWAS selected phenotypes (y-axis) enriched for protein coding genes identified with our SNEEP workflow. The x-axis shows the -log10 FDR corrected p-value of the disease enrichment analysis performed with the DisGeNET software.

References

Accili, D. and Arden, K. C. (2004). FoxOs at the crossroads of cellular metabolism, differentiation, and transformation. *Cell*, **117**(4), 421–426.

Alberts, R. *et al.* (2017). Genetic association analysis identifies variants associated with disease progression in primary sclerosing cholangitis. *Gut*, **67**(8), 1517–1524.

Alipanahi, B. *et al.* (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, **33**(8), 831–838.

Allen, R. J. *et al.* (2020). Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine*, **201**(5), 564–574.

Amariuta, T. *et al.* (2019). IMPACT: Genomic annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors. *The American Journal of Human Genetics*, **104**(5), 879–895.

Anders, S. *et al.* (2014). HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2), 166–169.

Anderson, W. F. *et al.* (1981). Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. *Nature*, **290**(5809), 754–758.

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

Anene-Nzelu, C. G. *et al.* (2020). Assigning distal genomic enhancers to cardiac disease–causing genes. *Circulation*, **142**(9), 910–912.

Arbogast, T. *et al.* (2016). Reciprocal effects on neurocognitive and metabolic phenotypes in mouse models of 16p11.2 deletion and duplication syndromes. *PLOS Genetics*, **12**(2), e1005709.

Arnold, C. D. *et al.* (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**(6123), 1074–1077.

Arnold, M. *et al.* (2014). SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics*, **31**(8), 1334–1336.

Avsec, Ž. *et al.* (2021a). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, **53**(3), 354–366.

Avsec, Ž. *et al.* (2021b). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, **18**(10), 1196–1203.

Baek, S. *et al.* (2011). Quantitative analysis of genome-wide chromatin remodeling. In *Methods in Molecular Biology*, pages 433–441. Humana Press.

Bailey, T. L. and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, **40**(17), e128.

Bain, G. (1994). E2a proteins are required for proper b cell development and initiation of immunoglobulin gene rearrangements. *Cell*, **79**(5), 885–892.

Bannister, A. J. *et al.* (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone h3 at active genes. *Journal of Biological Chemistry*, **280**(18), 17732–17736.

Baras, A. S. *et al.* (2015). Identification and validation of protein biomarkers of response to neoadjuvant platinum chemotherapy in muscle invasive urothelial carcinoma. *PLOS ONE*, **10**(7), 1–11.

Bardet, A. F. *et al.* (2013). Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics*, **29**(21), 2705–2713.

Barrangou, R. *et al.* (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**(5819), 1709–1712.

Barski, A. *et al.* (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, **129**(4), 823–837.

Baumgarten, N. *et al.* (2019). Improved linking of motifs to their TFs using domain information. *Bioinformatics*.

Baumgarten, N. *et al.* (2020). EpiRegio: analysis and retrieval of regulatory elements linked to genes. *Nucleic Acids Research*, **48**(W1), W193–W199.

Baumgarten, N. *et al.* (2021). Computational prediction of CRISPR-impaired noncoding regulatory regions. *Biological Chemistry*, **402**(8), 973–982.

Baumgarten, N. *et al.* (2023). A statistical approach to identify regulatory DNA variations. *bioRxiv*.

Beadle, G. W. and Tatum, E. L. (1941). Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of Sciences*, **27**(11), 499–506.

Beckstette, M. *et al.* (2006). Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, **7**(1).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Series B*, **57**, 289 – 300.

Berg, J. M. *et al.* (2013). *Biochemie*. Springer Spektrum, 7. edition.

Berg, O. G. and von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology*, **193**(4), 723–743.

Berger, M. F. *et al.* (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, **24**(11), 1429–1435.

Bhagwat, A. S. and Vakoc, C. R. (2015). Targeting transcription factors in cancer. *Trends in Cancer*, **1**(1), 53–65.

Bischoff, F. C. *et al.* (2017). Identification and functional characterization of hypoxia-induced endoplasmic reticulum stress regulating lncRNA (HypERlnc) in pericytes. *Circulation Research*, **121**(4), 368–375.

Blackwood, E. M. and Kadonaga, J. T. (1998). Going the distance: A current view of enhancer action. *Science*, **281**(5373), 60–63.

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Frontiers in Genetics*, **7**.

Boyle, A. P. *et al.* (2008). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**(21), 2537–2538.

Boytsov, A. *et al.* (2022). Positional weight matrices have sufficient prediction power for analysis of noncoding variants [version 3; peer review: 2 approved]. *F1000Research*, **11**, 33.

Bray, N. L. *et al.* (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5), 525–527.

Broekema, R. V. *et al.* (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biology*, **10**(1), 190221.

Brown, A. A. *et al.* (2017). Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nature Genetics*, **49**(12), 1747–1751.

Buckley, A. *et al.* (2019). Characterisation of an Isogenic Model of Cisplatin Resistance in Oesophageal Adenocarcinoma Cells. *Pharmaceuticals*, **12**(1), 33.

Budhavarapu, V. N. *et al.* (2013). How is epigenetic information maintained through DNA replication? *Epigenetics and Chromatin*, **6**(1).

Bujold, D. *et al.* (2016). The international human epigenome consortium data portal. *Cell Systems*, **3**(5), 496–499.e2.

Buniello, A. *et al.* (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, **47**(D1), D1005–D1012.

Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm.

Cai, Q. *et al.* (2020). IGFBP-3/IGFBP-3 Receptor System as an Anti-Tumor and Anti-Metastatic Signaling in Cancer. *Cells*, **9**(5), 1261.

Cai, Z. *et al.* (2019). RAEdb: a database of enhancers identified by high-throughput reporter assays. *Database*, **2019**.

Campbell, N. A. and Reece, J. B. (2009). *Biologie*. Pearson, 8. edition.

Castro-Mondragon, J. A. *et al.* (2021). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **50**(D1), D165–D173.

Cattaneo, P. *et al.* (2022). DOT1l regulates chamber-specific transcriptional networks during cardiogenesis and mediates postnatal cell cycle withdrawal. *Nature Communications*, **13**(1).

Chan, H. P. *et al.* (2010). Importance sampling of word patterns in DNA and protein sequences. *Journal of Computational Biology*, **17**(12), 1697–1709.

Chen, F. X. *et al.* (2017). PAF1 regulation of promoter-proximal pause release via enhancer activation. *Science*, **357**(6357), 1294–1298.

Chen, J. *et al.* (2004). Detection of functional DNA motifs via statistical overrepresentation. *Nucleic Acids Research*, **32**(4), 1372–1381.

Chen, K. *et al.* (2012). DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research*, **23**(2), 341–351.

Chen, K.-B. and Zhang, Y. (2010). A varying threshold method for ChIP peak-calling using multiple sources of information. *Bioinformatics*, **26**(18), i504–i510.

Chen, L. *et al.* (2022). Exploiting deep transfer learning for the prediction of functional non-coding variants using genomic sequence. *Bioinformatics*, **38**(12), 3164–3172.

Cho, S. W. *et al.* (2018). Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. *Cell*, **173**(6), 1398–1412.e22.

Choi, C.-H. (2005). Abc transporters as multidrug resistance mechanisms and the development of chemosensitizers for their reversal. *Cancer Cell International*, **5**(1), 30.

Conesa, A. *et al.* (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, **17**(1).

Creyghton, M. P. *et al.* (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, **107**(50), 21931–21936.

de Almeida, B. P. *et al.* (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, **54**(5), 613–624.

Deisenroth, M. P. *et al.* (2020). *Mathematics for Machine Learning*. Cambridge University Press.

Dekking, F. (2005). *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer Texts in Statistics. Springer.

DeLong, E. R. *et al.* (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, **44**(3), 837.

Deplancke, B. *et al.* (2016). The genetics of transcription factor DNA binding variation. *Cell*, **166**(3), 538–554.

Dey, K. K. *et al.* (2022). SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell Genomics*, **2**(7), 100145.

Dieck, C. L. and Ferrando, A. (2019). Genetics and mechanisms of NT5C2-driven chemotherapy resistance in relapsed ALL. *Blood*, **133**(21), 2263–2268.

Dobin, A. *et al.* (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1), 15–21.

Dukler, N. *et al.* (2017). Is a super-enhancer greater than the sum of its parts? *Nature Genetics*, **49**(1), 2–3.

Dupont, C. *et al.* (2009). Epigenetics: Definition, mechanisms and clinical perspective. *Seminars in Reproductive Medicine*, **27**(05), 351–357.

Dvash, E. *et al.* (2015). Leukotriene C4 is the major trigger of stress-induced oxidative DNA damage. *Nature Communications*, **6**(1).

Dynlacht, B. D. (1997). Regulation of transcription by proteins that control the cell cycle. *Nature*, **389**(6647), 149–152.

Eichler, E. E. *et al.* (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**(6), 446–450.

El-Gebali, S. *et al.* (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, **47**(D1), D427–D432.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, **28**(8), 817–825.

Essletzbichler, P. *et al.* (2014). Megabase-scale deletion using CRISPR/cas9 to generate a fully haploid human cell line. *Genome Research*, **24**(12), 2059–2065.

Fang, F. *et al.* (2013). Early growth response 3 (egr-3) is induced by transforming growth factor-β and regulates fibrogenic responses. *The American Journal of Pathology*, **183**(4), 1197–1208.

Ferragina, P. and Manzini, G. (2001). An experimental study of an opportunistic index. In *ACM-SIAM Symposium on Discrete Algorithms*.

Fisher, R. A. (1934). *Statistical Methods for Research Workers (4th ed.)*. Edinburgh: Oliver and Boyd.

Fishilevich, S. *et al.* (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, **2017**.

Flavahan, W. A. *et al.* (2015). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**(7584), 110–114.

Flitsch, L. J. *et al.* (2020). Transcription factor-based fate specification and forward programming for neural regeneration. *Frontiers in Cellular Neuroscience*, **14**.

Forbes, C. *et al.* (2010). *Statistical Distributions*. John Wiley & Sons, Inc.

Fornes, O. *et al.* (2019). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*.

Francis, N. J. and Sihou, D. (2021). Inheritance of histone (h3/h4): A binary choice? *Trends in Biochemical Sciences*, **46**(1), 5–14.

Frankish, A. *et al.* (2020). GENCODE 2021. *Nucleic Acids Research*, **49**(D1), D916–D923.

Fry, A. *et al.* (2017). Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *American Journal of Epidemiology*, **186**(9), 1026–1034.

Fu, Y. *et al.* (2004). MotifViz: an analysis and visualization tool for motif discovery. *Nucleic Acids Research*, **32**(Web Server), W420–W423.

Fulco, C. P. *et al.* (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, **51**(12), 1664–1669.

Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, **13**, 840–852.

Furlong, E. E. M. and Levine, M. (2018). Developmental enhancers and chromosome topology. *Science*, **361**(6409), 1341–1345.

Galindo-Pumariño, C. *et al.* (2022). SNAI1-expressing fibroblasts and derived-extracellular matrix as mediators of drug resistance in colorectal cancer patients. *Toxicology and Applied Pharmacology*, **450**, 116171.

Gao, T. and Qian, J. (2019a). EAGLE: An algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. *PLOS Computational Biology*, **15**(10), e1007436.

Gao, T. and Qian, J. (2019b). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research.*

Gao, Y. *et al.* (2020). Identification of gene modules associated with survival of diffuse large B-cell lymphoma treated with CHOP-based chemotherapy. *The Pharmacogenomics Journal*, **20**(5), 705–716.

Gao, Y. *et al.* (2023). Role of transcription factors in apoptotic cells clearance. *Frontiers in Cell and Developmental Biology*, **11**.

Gaspar-Maia, A. *et al.* (2009). Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature*, **460**(7257), 863–868.

Gasperini, M. *et al.* (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, **21**(5), 292–310.

Gazal, S. *et al.* (2022). Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics*, **54**(6), 827–836.

Gerstein, M. B. *et al.* (2007). What is a gene, post-ENCODE? history and updated definition. *Genome Research*, **17**(6), 669–681.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, **13**(2), 135–145.

Gilfillan, G. D. *et al.* (2012). Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics*, **13**(1), 645.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, **81**(25), 2340–2361.

Glaser, S. F. *et al.* (2022). Circular RNA circPLOD2 regulates pericyte function by targeting the transcription factor KLF4.

Grant, C. E. *et al.* (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**(7), 1017–1018.

Grau, J. *et al.* (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, **31**(15), 2595–2597.

Gray, P. A. *et al.* (2004). Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science*, **306**(5705), 2255–2257.

Greenberg, M. V. C. and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, **20**(10), 590–607.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. The MIT Press.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature*, **550**(7675), 204–213.

Gurha, P. (2019). Noncoding RNAs in cardiovascular diseases. *Current Opinion in Cardiology*, **34**(3), 241–245.

Hai, T. *et al.* (1999). Atf3 and stress responses. *Gene Expression*, **7**(4-6), 321–335.

Hargreaves, D. C. and Crabtree, G. R. (2011). ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell Research*, **21**(3), 396–420.

Harrison, R. L. *et al.* (2010). Introduction to monte carlo simulation. In *AIP Conference Proceedings*. AIP.

He, H. *et al.* (2021). HFBD: a biomarker knowledge database for heart failure heterogeneity and personalized applications. *Bioinformatics*, **37**(23), 4534–4539.

He, H. H. *et al.* (2010). Nucleosome dynamics define transcriptional enhancers. *Nature Genetics*, **42**(4), 343–347.

He, Q. *et al.* (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, **33**(4), 395–401.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. *Biometrika*, **105**(1), 239–246.

Hecker, D. *et al.* (2023). The adapted activity-by-contact model for enhancer–gene assignment and its application to single-cell data. *Bioinformatics*, **39**(2).

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

Heinz, S. *et al.* (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, **16**(3), 144–154.

Hnisz, D. *et al.* (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**(6280), 1454–1458.

Hormozdiari, F. *et al.* (2016). Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics*, **99**(6), 1245–1260.

Hu, L. *et al.* (2019). Targeting TR4 nuclear receptor with antagonist bexarotene increases docetaxel sensitivity to better suppress the metastatic castration-resistant prostate cancer progression. *Oncogene*, **39**(9), 1891–1903.

Hume, M. A. *et al.* (2015). UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, **43**(D1), D117–D122.

Hunter, S. *et al.* (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research*, **37**(Database), D211–D215.

Ibrahim, M. M. *et al.* (2014). JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, **31**(1), 48–55.

Ignatieva, E. V. *et al.* (2015). Human genes encoding transcription factors and chromatin-modifying proteins have low levels of promoter polymorphism: A study of 1000 genomes project data. *International Journal of Genomics*, **2015**, 1–15.

Indermaur, M. D. *et al.* (2010). Genomic-directed targeted therapy increases endometrial cancer cell sensitivity to doxorubicin. *American Journal of Obstetrics and Gynecology*, **203**(2), 158.e1 – 158.e40.

Islam, Z. *et al.* (2021). Transcription factors: The fulcrum between cell development and carcinogenesis. *Frontiers in Oncology*, **11**.

James, G. *et al.* (2013). *An Introduction to Statistical Learning*. Springer New York.

Janky, R. *et al.* (2014). iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Computational Biology*, **7**.

Jeck, W. R. *et al.* (2012). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**(2), 141–157.

Jeong, Y. *et al.* (2008). Regulation of a remote shh forebrain enhancer by the six3 homeoprotein. *Nature Genetics*, **40**(11), 1348–1353.

Jiang, S. and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data. *Briefings in Functional Genomics*, **17**(2), 104–115.

Jiao, X. *et al.* (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**(13), 1805–1806.

Jinek, M. *et al.* (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**(6096), 816–821.

Johannsen, W. (1909). Gustav Fischer.

Jolma, A. *et al.* (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, **20**(6), 861–873.

K. Blitzstein, J. and Hwang, J. (2015). *Introduction to Probability Second Edition*. CRC Press.

Kasowski, M. *et al.* (2010). Variation in transcription factor binding among humans. *Science*, **328**(5975), 232–235.

Keepers, B. *et al.* (2020). What's in a cardiomyocyte – and how do we make one through reprogramming? *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1867**(3), 118464.

Kehl, T. *et al.* (2018). REGGAE: a novel approach for the identification of key transcriptional regulators. *Bioinformatics*, **34**(20), 3503–3510.

Keilwagen, J. and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Research*, **43**(18), e119–e119.

Kelley, D. R. *et al.* (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, **26**(7), 990–999.

Kelley, D. R. *et al.* (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, **28**(5), 739–750.

Khan, A. *et al.* (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, **46**(D1), D260–D266.

Kheradpour, P. and Kellis, M. (2013). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, **42**(5), 2976–2987.

Kim, S. and Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell*, **83**(3), 373–392.

Kim, S. S. *et al.* (2019). Genes with high network connectivity are enriched for disease heritability. *The American Journal of Human Genetics*, **104**(5), 896–913.

Klann, T. S. *et al.* (2017). CRISPR–cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature Biotechnology*, **35**(6), 561–568.

Klein, D. C. and Hainer, S. J. (2019). Genomic methods in profiling DNA accessibility and factor localization. *Chromosome Research*, **28**(1), 69–85.

Kojima, Y. *et al.* (2021). GATA transcription factors, SOX17 and TFAP2c, drive the human germ-cell specification program. *Life Science Alliance*, **4**(5), e202000974.

Koohy, H. *et al.* (2014). A comparison of peak callers used for DNase-seq data. *PLoS ONE*, **9**(5), e96303.

Kotz, S. *et al.* (2001). *The Laplace Distribution and Generalizations*. Birkhäuser Boston.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, **128**(4), 693–705.

Kroese, D. P. *et al.* (2014). Why the monte carlo method is so important today. *WIREs Computational Statistics*, **6**(6), 386–392.

Kukurba, K. R. and Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, **2015**(11), pdb.top084970.

Kulakovskiy, I. V. *et al.* (2017). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Research*, **46**(D1), D252–D259.

Kumasaka, N. *et al.* (2015). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, **48**(2), 206–213.

Kundaje, A. *et al.* (2015). Integrative analysis of 111 reference human epigenomes.

Kuttippurathu, L. *et al.* (2010). CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. *Bioinformatics*, **27**(5), 715–717.

Lamb, T. D. (2015). Why rods and cones? *Eye*, **30**(2), 179–185.

Lambert, S. A. *et al.* (2018). The human transcription factors. *Cell*, **172**(4), 650–665.

Landrum, M. J. *et al.* (2017). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, **46**(D1), D1062–D1067.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, **9**(4), 357–359.

Lee, B.-K. *et al.* (2011). Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Research*, **22**(1), 9–24.

Lee, D. *et al.* (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, **47**(8), 955–961.

Lehman, I. and Kaguni, L. S. (1989). DNA ploymerase alpha. *The Journal of Biological Chemistry*, **264**(8), 4265–4268.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Lettice, L. A. (2003). A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, **12**(14), 1725–1735.

Letunic, I. *et al.* (2015). SMART: recent updates, new developments and status in 2015. *Nucleic Acids Research*, **43**(D1), D257–D260.

Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, **102**(14), 4936–4942.

Levings, P. P. and Bungert, J. (2002). The human β-globin locus control region. *European Journal of Biochemistry*, **269**(6), 1589–1599.

Li, G. *et al.* (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**(1-2), 84–98.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Li, L. and Wunderlich, Z. (2017). An enhancer's length and composition are shaped by its regulatory task. *Frontiers in Genetics*, **8**.

Li, M. J. *et al.* (2015). GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Research*, **44**(D1), D869–D876.

Li, M.-X. *et al.* (2011). GATES: A rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics*, **88**(3), 283–293.

Liu, H. *et al.* (2020). ANXA7 promotes the cell cycle, proliferation and cell adhesion-mediated drug resistance of multiple myeloma cells by up-regulating CDC5L. *Aging*, **12**(11), 11100–11115.

Liu, J. Z. *et al.* (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, **87**(1), 139–145.

Liu, N. *et al.* (2022). Circular RNA circTmem241 drives group III innate lymphoid cell differentiation via initiation of elk3 transcription. *Nature Communications*, **13**(1).

Liu, Y. *et al.* (2017). Transcriptional landscape of the human cell cycle. *Proceedings of the National Academy of Sciences*, **114**(13), 3473–3478.

Loewe, L. and Hill, W. G. (2010). The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**(1544), 1153–1167.

Long, H. K. *et al.* (2020). Loss of extreme long-range enhancers in human neural crest drives a craniofacial disorder. *Cell Stem Cell*, **27**(5), 765–783.e14.

Love, M. I. *et al.* (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**(12).

Luo, Y. *et al.* (2019). New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Research*, **48**(D1), D882–D889.

Luscombe, N. M. *et al.* (2000). An overview of the structures of protein-DNA complexes. *Genome Biology*, **1**(1), reviews001.1001.37.

Lyon, M. F. (1961). Gene action in the x-chromosome of the mouse (mus musculus l.). *Nature*, **190**(4773), 372–373.

Ma, J. (2005). Crossing the line between activation and repression. *Trends in Genetics*, **21**(1), 54–59.

Ma, W. *et al.* (2014). Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nature Protocols*, **9**(6), 1428–1450.

Machiela, M. J. and Chanock, S. J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, **31**(21), 3555–3557.

Macintyre, G. *et al.* (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **26**(18), i524–i530.

Manke, T. *et al.* (2010). Quantifying the effect of sequence variation on regulatory interactions. *Human Mutation*, **31**(4), 477–483.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, **18**(1), 50–60.

Maricque, B. B. *et al.* (2018). A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nature Biotechnology*, **37**(1), 90–95.

Martin, V. *et al.* (2019). QBiC-pred: quantitative predictions of transcription factor binding changes due to sequence variants. *Nucleic Acids Research*, **47**(W1), W127–W135.

Mathelier, A. and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS Computational Biology*, **9**(9), e1003214.

Mattick, J. S. *et al.* (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature Reviews Molecular Cell Biology*.

Matys, V. *et al.* (2006). TRANSFACő and its module TRANSCompelő: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, **34**, D108 – D110.

Maurice, D. *et al.* (2018). ERK signaling controls innate-like CD8+ t cell differentiation via the ELK4 (SAP-1) and ELK1 transcription factors. *The Journal of Immunology*, **201**(6), 1681–1691.

McDonald, J. (2014). *Handbook of Biological Statistics*. Sparky House Publishing, Balitmore, Maryland, third edition.

McLeay, R. C. and Bailey, T. L. (2010). Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**(1).

Memczak, S. *et al.* (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**(7441), 333–338.

Mengstie, M. A. and Wondimu, B. Z. (2021). Mechanism and applications of CRISPR/cas-9-mediated genome editing. *Biologics: Targets and Therapy*, **Volume 15**, 353–361.

Metropolis, N. *et al.* (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Mikhaylichenko, O. *et al.* (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes and Development*, **32**(1), 42–57.

Monahan, K. *et al.* (2019). LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature*, **565**(7740), 448–453.

Moon, J.-S. *et al.* (2011). Krüppel-like factor 4 (KLF4) activates the transcription of the gene for the platelet isoform of phosphofructokinase (PFKP) in breast cancer. *Journal of Biological Chemistry*, **286**(27), 23808–23816.

Nadler, Y. *et al.* (2010). Growth factor receptor-bound protein-7 (grb7) as a prognostic marker and therapeutic target in breast cancer. *Annals of Oncology*, **21**(3), 466–473.

Nagai, A. *et al.* (2017). Overview of the BioBank japan project: Study design and profile. *Journal of Epidemiology*, **27**(3), S2–S8.

Nakato, R. and Shirahige, K. (2016). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in Bioinformatics*, page bbw023.

Nayak, B. and Hazra, A. (2011). How to choose the right statistical test? *Indian Journal of Ophthalmology*, **59**(2), 85.

Nelson, C. P. *et al.* (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genetics*, **49**(9), 1385–1391.

Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**(1620), 20120362.

Nobori, K. *et al.* (2002). ATF3 inhibits doxorubicin-induced apoptosis in cardiac myocytes: a novel cardioprotective role of ATF3. *J. Mol. Cell. Cardiol.*, **34**(10), 1387–1397.

Nordheim, A. and Knippers, R. (2018). *Molekular Genetik*. Thieme, 11. edition.

O'Neill, L. (2003). Immunoprecipitation of native chromatin: NChIP. *Methods*, **31**(1), 76–82.

Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends in Biochemical Sciences*, **25**(3), 99–104.

Pape, U. J. *et al.* (2008). Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, **24**(3), 350–357.

Park, E. J. *et al.* (2012). Doxorubicin induces cytotoxicity through upregulation of pERK-dependent ATF3. *PLoS ONE*, **7**(9), e44990.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**(10), 669–680.

Patro, R. *et al.* (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, **14**(4), 417–419.

Pickar-Oliver, A. and Gersbach, C. A. (2019). The next generation of CRISPR–cas technologies and applications. *Nature Reviews Molecular Cell Biology*, **20**(8), 490–507.

Pinero, J. *et al.* (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**(0), bav028–bav028.

Piñero, J. *et al.* (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*.

Portin, P. and Wilkins, A. (2017). The evolving definition of the term "gene". *Genetics*, **205**(4), 1353–1364.

Potier, D. *et al.* (2015). i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Research*, **43**(W1), W57–W64.

Purcell, S. *et al.* (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**(3), 559–575.

Qu, K. *et al.* (2015a). Individuality and variation of personal regulomes in primary human t cells. *Cell Systems*, **1**(1), 51–61.

Qu, S. *et al.* (2015b). Circular RNA: A new star of noncoding RNAs. *Cancer Letters*, **365**(2), 141–148.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

Rada-Iglesias, A. *et al.* (2010). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**(7333), 279–283.

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

Rao, S. S. *et al.* (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**(7), 1665–1680.

Rashid, N. U. *et al.* (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*, **12**(7), R67.

Rath, A. *et al.* (2012). Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, **33**(5), 803–808.

Raudvere, U. *et al.* (2019). g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, **47**(W1), W191–W198.

Rawofi, L. *et al.* (2017). Genome-wide association study of pigmentary traits (skin and iris color) in individuals of east asian ancestry. *PeerJ*, **5**, e3951.

Rehm, H. L. *et al.* (2015). ClinGen — the clinical genome resource. *New England Journal of Medicine*, **372**(23), 2235–2242.

Reiter, F. *et al.* (2017). Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics & Development*, **43**, 73–81.

Richmond, T. J. and Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature*, **423**(6936), 145–150.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(3), 507–554.

Riggins, R. B. *et al.* (2010). Orphan nuclear receptors in breast cancer pathogenesis and therapeutic response. *Endocrine-Related Cancer*, **17**(3), R213–R231.

Robin, X. *et al.* (2011). pROC: an open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**(1).

Rochette, L. *et al.* (2020). Mitochondrial SLC25 Carriers: Novel Targets for Cancer Therapy. *Molecules*, **25**(10), 2417.

Roider, H. G. *et al.* (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**(2), 134–141.

Roider, H. G. *et al.* (2009). PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, **25**(4), 435–442.

Roy, A. L. and Conroy, R. S. (2018). Toward mapping the human body at a cellular resolution. *Molecular Biology of the Cell*, **29**(15), 1779–1785.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, **10**(3), e0118432.

Salzberg, S. L. (2018). Open questions: How many genes do we have? *BMC Biology*, **16**(1).

Salzman, J. *et al.* (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE*, **7**(2), e30733.

Santa, F. D. *et al.* (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biology*, **8**(5), e1000384.

Sanyal, A. *et al.* (2012). The long-range interaction landscape of gene promoters. *Nature*, **489**(7414), 109–113.

Schmidt, F. *et al.* (2021). Integrative analysis of epigenetics data identifies gene-specific regulatory elements. *Nucleic Acids Research*, **49**(18), 10397–10418.

Schneider, T. D. and Stephens, R. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, **18**(20), 6097–6100.

Schröder, Adrian, J. *et al.* (2010). Predicting DNA-Binding Specificities of Eukaryotic Transcription Factors. *PLoS ONE*, **5**(11), 1–15.

Schulze, R. J. *et al.* (2019). The cell biology of the hepatocyte: A membrane trafficking machine. *Journal of Cell Biology*, **218**(7), 2096–2112.

Schuster, S. L. and Hsieh, A. C. (2019). The untranslated regions of mRNAs in cancer. *Trends in Cancer*, **5**(4), 245–262.

Schwartzentruber, J. *et al.* (2012). Driver mutations in histone h3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature*, **482**(7384), 226–231.

Shen, J. *et al.* (2020). PFKP is highly expressed in lung cancer and regulates glucose metabolism. *Cellular Oncology*, **43**(4), 617–629.

Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**(1), 308–311.

Shi, W. *et al.* (2016). Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Research*, page gkw691.

Simon, I. *et al.* (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**(6), 697–708.

Slack, F. J. and Chinnaiyan, A. M. (2019). The role of non-coding RNAs in oncology. *Cell*, **179**(5), 1033–1055.

Spallarossa, P. *et al.* (2006). Matrix metalloproteinase-2 and -9 are induced differently by doxorubicin in H9c2 cells: The role of MAP kinases and NAD(P)H oxidase. *Cardiovasc. Res.*, **69**(3), 736–745.

Spicuglia, S. and Vanhille, L. (2012). Chromatin signatures of active enhancers. *Nucleus*, **3**(2), 126–131.

Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, **13**(9), 613–626.

Stark, R. *et al.* (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, **20**(11), 631–656.

Steinhaus, R. *et al.* (2022). FABIAN-variant: predicting the effects of DNA variants on transcription factor binding. *Nucleic Acids Research*, **50**(W1), W322–W329.

Stenzel, O. *et al.* (2012). A new approach to model-based simulation of disordered polymer blend solar cells. *Advanced Functional Materials*, **22**(6), 1236–1244.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.

Stunnenberg, H. G. *et al.* (2016). The international human epigenome consortium: A blueprint for scientific collaboration and discovery. *Cell*, **167**(5), 1145–1149.

Sun, K. *et al.* (2017). Ncoa5 promotes proliferation, migration and invasion of colorectal cancer cells via activation of PI3k/AKT pathway. *Oncotarget*, **8**(64), 107932–107946.

Szklarczyk, D. *et al.* (2020). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, **49**(D1), D605–D612.

Tan, K. *et al.* (2005). Making connections between novel transcription factors and their DNA motifs. *Genome Research*, **15**, 312–320.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319), 1061–1073.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489 7414**, 57–74.

The ENCODE Project Consortium., Snyder, M. G. T. e. a. (2020). Perspectives on ENCODE. *Nature*, **583**(7818), 693–698.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **45**(D1), D158–D169.

Thiedmann, R. *et al.* (2011). Stochastic simulation model for the 3d morphology of composite materials in li–ion batteries. *Computational Materials Science*, **50**(12), 3365–3376.

Tombor, L. S. *et al.* (2021). Single cell sequencing reveals endothelial plasticity with transient mesenchymal activation after myocardial infarction. *Nature Communications*, **12**(1).

Tompa, M. *et al.* (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, **23**(1), 137–144.

Tran, N. T. L. and Huang, C.-H. (2014). A survey of motif finding web tools for detecting binding site motifs in ChIP-seq data. *Biology Direct*, **9**(1), 4.

Trapnell, C. *et al.* (2009). TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, **25**(9), 1105–1111.

Trapnell, C. *et al.* (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nature Protocols*, **7**(3), 562–578.

Tsiomita, S. *et al.* (2022). ETS2 repressor factor (ERF) is involved in t lymphocyte maturation acting as regulator of thymocyte lineage commitment. *Journal of Leukocyte Biology*, **112**(4), 641–657.

Tsompana, M. and Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics and Chromatin*, **7**(1).

Uffelmann, E. *et al.* (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, **1**(1).

van der Valk, R. J. *et al.* (2014). A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Human Molecular Genetics*, **24**(4), 1155–1168.

van Heeringen, S. J. and Veenstra, G. J. C. (2010). GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics*, **27**(2), 270–271.

Vandiedonck, C. (2018). Genetic association of molecular traits: A help to identify causative variants in complex diseases. *Clinical Genetics*, **93**(3), 520–532.

Vaquerizas, J. M. *et al.* (2009). A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, **10**(4), 252–263.

Verheul, T. C. J. *et al.* (2020). The why of YY1: Mechanisms of transcriptional regulation by yin yang 1. *Frontiers in Cell and Developmental Biology*, **8**.

Virtanen, P. *et al.* (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, **17**(3), 261–272.

Visel, A. *et al.* (2007). VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Research*, **35**(Database), D88–D92.

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

Wagner, G. P. *et al.* (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, **131**(4), 281–285.

Walczynski, J. *et al.* (2013). Sensitisation of c-MYC-induced b-lymphoma cells to apoptosis by ATF2. *Oncogene*, **33**(8), 1027–1036.

Wang, G. *et al.* (2020). Slc25a10 performs an oncogenic role in human osteosarcoma. *Oncol Lett*, **20**(4), 2.

Wang, J. *et al.* (2013). BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics*, **29**(4), 492–493.

Wang, J. *et al.* (2018a). HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Research*, **47**(D1), D106–D112.

Wang, X. *et al.* (2016). COUP-TFII suppresses colorectal carcinoma resistance to doxorubicin involving inhibition of epithelial-mesenchymal transition. *American journal of translational research*, **8**(9), 3921–3929.

Wang, X. *et al.* (2019). Fatty acid receptor GPR120 promotes breast cancer chemoresistance by upregulating ABC transporters expression and fatty acid synthesis. *EBioMedicine*, **40**, 251–262.

Wang, Y. *et al.* (2018b). AP-1 confers resistance to anti-cancer therapy by activating XIAP. *Oncotarget*, **9**(18), 14124–14137.

Wang, Z. *et al.* (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1), 57–63.

Wang, Z. *et al.* (2017). HEDD: Human enhancer disease database. *Nucleic Acids Research*, **46**(D1), D113–D120.

Wasserman, L. (2010). *All of statistics : a concise course in statistical inference*. Springer, New York.

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, **5**(4), 276–287.

Wegner, M. *et al.* (2019). Circular synthesized CRISPR/cas gRNAs for functional interrogations in the coding and noncoding genome. *eLife*, **8**.

Weingarten-Gabbay, S. *et al.* (2019). Systematic interrogation of human promoters. *Genome Research*, **29**(2), 171–183.

Wienert, B. *et al.* (2015). Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. *Nature Communications*, **6**(1).

Wingender, E. *et al.* (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, **41**(D1), D165–D170.

Wingender, E. *et al.* (2017). TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research*, **46**(D1), D343–D347.

Wu, D. *et al.* (2016). The emerging roles of orphan nuclear receptors in prostate cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, **1866**(1), 23–36.

Wyman, C. and Kanaar, R. (2006). DNA double-strand break repair: All's well that ends well. *Annual Review of Genetics*, **40**(1), 363–383.

Xu, S. *et al.* (2014). Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. In *Methods in Molecular Biology*, pages 97–111. Springer New York.

Yan, J. *et al.* (2021). Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, **591**(7848), 147–151.

Yoshida, H. *et al.* (2019). The cis-regulatory atlas of the mouse immune system. *Cell*, **176**(4), 897–912.e20.

Zamanighomi, M. *et al.* (2017). Predicting transcription factor binding motifs from DNA-binding domains, chromatin accessibility and gene expression data. *Nucleic Acids Research*, **45**(10), 5666–5677.

Zar, J. (2010). *Biostatistical Analysis*. Pearson Educatiom, fifth edition.

Zaret, K. S. (2018). Pioneering the chromatin landscape. *Nature Genetics*, **50**(2), 167–169.

Zeitlinger, J. (2020). Seven myths of how transcription factors read the cis-regulatory code. *Current Opinion in Systems Biology*, **23**, 22–31.

Zeng, H. *et al.* (2015). GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, **32**(4), 490–496.

Zhang, F. and Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, **24**(R1), R102–R110.

Zhang, G. *et al.* (2017). DiseaseEnhancer: a resource of human disease-associated enhancer catalog. *Nucleic Acids Research*, **46**(D1), D78–D84.

Zhang, Y. *et al.* (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biology*, **9**(9).

Zhao, J. *et al.* (2017). Quantifying the impact of non-coding variants on transcription factor-DNA binding. In *Lecture Notes in Computer Science*, pages 336–352. Springer International Publishing.

Zhao, L. *et al.* (2019). Nuclear receptors: recent drug discovery for cancer therapies. *Endocrine Reviews*.

Zhao, Y. *et al.* (2012). Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**(3), 781–790.

Zheng, X. *et al.* (2014). A rare duplication on chromosome 16p11.2 is identified in patients with psychosis in alzheimer's disease. *PLoS ONE*, **9**(11), e111462.

Zheng, Y. *et al.* (2013). Epigenetic downregulation of runx3 by dna methylation induces docetaxel chemoresistance in human lung adenocarcinoma cells by activation of the akt pathway. *The International Journal of Biochemistry and Cell Biology*, **45**(11), 2369 – 2378.

Zhu, C. *et al.* (2023). Cvd-associated snps with regulatory potential drive pathologic non-coding rna expression. *bioRxiv*.

Zhu, X. and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature Communications*, **9**(1).

Zuo, C. *et al.* (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics*, **31**(20), 3353–3355.

Zusammenfassung

Eine zentrale Fragestellung der molekularen Genetik befasst sich damit herauszufinden, wie ein Organismus die genetischen und epigenetischen Informationen in einen Phänotypen umsetzt. Ausgehend von einer Zygote entwickelt sich der menschliche Körper, welcher aus Billionen von Zellen besteht, die sich in circa 200 verschiedene Zelltypen unterteilen lassen (Roy and Conroy, 2018). Jeder Zelltyp erfüllt spezifische Aufgaben. Ein Beispiel sind Kardiomyozyten, die für die Kontraktion des Herzens verantwortlich sind (Keepers *et al.*, 2020). Hepatozyten, der am häufigsten vorkommende Zelltype in der Leber, ist unter Anderem daran beteiligt das Blut zu entgiften (Schulze *et al.*, 2019). Weitere Beispiele sind T-Lymphozyten, die in der Immunantwort des Körpers durch die Etablierung der zellulären Immunität eine wichtige Rolle spielen (Campbell and Reece, 2009, p. 1290), oder Stäbchen und Zapfen, Zellen des Auges, die zum menschlichen Sehvermögen beitragen (Lamb, 2015).
Wie etablieren Zellen ihre unterschiedlichen Funktionalitäten, obwohl sie alle auf der gleichen DNA basieren? Die Expression der Gene unterscheidet sich von Zelle zu Zelle, nicht nur darin wie stark ein Gen transkribiert ist, sondern auch, welche Gene exprimiert sind. Genexpression wird von Transkriptionsfaktoren (TFs) reguliert, Proteinen, die die Fähigkeit besitzen spezifische DNA Muster, bekannt als Motive, zu erkennen und dadurch an die DNA binden (Lambert *et al.*, 2018). Diese Motive treten besonders häufig in regularischen Elementen (REMs) wie Promotoren und Enhancern auf (Gasperini *et al.*, 2020). Aktive REMs zeichnen sich dadurch aus, dass sie von TFs und ihren Co-Faktoren gebunden werden und die Genexpression beeinflussen. Promotoren sind genomische Regionen die stromaufwärts vom Transkriptionsstartpunkt liegen und die Transkription initiieren. Im Gegensatz dazu unterstützen Enhancer die Genexpression, indem sie über 3D-Schleifen mit dem Promotor interagieren, selbst wenn sie tausende von Basenpaaren entfernt liegen.
Mit dem Aufkommen von Next-Generation-Sequenzierung (NGS) wurde er-

wartet, dass der sogenannte cis-regulatorische Code, der die Zusammenhänge zwischen der DNA Sequenz und der Genexpression beschreibt, schnell entschlüsselt wird (Zeitlinger, 2020; Kim and Wysocka, 2023). Es hat sich jedoch herausgestellt, dass dies nicht so einfach ist wie beim genetischen Code, der es ermöglicht, anhand der DNA Sequenz die Aminosäuresequenz zu bestimmen. Ein möglicher Grund dafür ist, dass REMs stark zelltyp-spezifisch sind, wobei einige von ihnen nur zu bestimmten Zeitpunkten aktiv sind, wie zum Beispiel während der embryonalen Entwicklung (Levine and Davidson, 2005). Des Weiteren ist nicht klar, wo in der DNA Sequenz die REMs lokalisiert sind, die ein Gen regulieren. Für einige REMs is bekannt, dass sie Megabasenpaare vom Promotor entfernt liegen mit dem sie kommunizieren (Long *et al.*, 2020; Lettice, 2003) oder sie interagieren sogar zwischen verschiedenen Chromosomen (Monahan *et al.*, 2019). Zusätzlich ist die quantitativ präzise Regulierung der Gene äußerst wichtig, denn schon moderate Veränderung der Genexpressionsstärke kann zu Krankheiten wie zum Beispiel Krebs (Flavahan *et al.*, 2015; Cho *et al.*, 2018) oder Herz-Kreislauf-Erkrankungen führen (Yoshida *et al.*, 2019). Aus diesen Gründen ist es ein zentrales Ziel der Genetik zu verstehen, wie Gene reguliert werden. In dieser Arbeit möchten wir dazu beitragen, die Genregulation besser zu verstehen und bearbeiten drei verschiedene Themen:

Eine wesentliche Rolle bei der Regulation der Gene spielen Transkriptionsfaktoren (TFs), da sie in der Lage sind direkt mit der DNA zu interagieren. Ein erster Schritt zur Erforschung der Auswirkungen eines TFs auf die Genregulation besteht häufig darin, die genomischen Regionen zu identifizieren, die vom TF gebunden werden. Dafür muss man die Bindepräferenzen des TFs kennen, die üblicherweise in TF-Bindungsmotiven zusammengefasst werden. Obwohl für viele TFs ein Bindungsmotiv experimentell validiert ist, gibt es immer noch eine große Anzahl von TFs, für die kein Bindungsmotiv bekannt ist. Aus diesem Grund gibt es einige Tools, die Bindungsmotive mit TFs verknüpfen. In dieser Arbeit stellen wir eine Methode, genannt Massif, vor, die die Vorhersage von solchen Tools verbessert, indem sie die DNA Bindungsdomäne (DBD) des zu untersuchenden TFs mit einbezieht (siehe Kapitel 3).
Wir haben beobachtet, dass TFs mit derselben DBD dazu tendieren ähnliche Bindepräferenzen, und damit auch ähnliche Motive aufzuweisen. Aus diesem Grund nehmen wir an, dass ein mögliches Motiv für einen TF mit größerer Wahrscheinlichkeit das Richtige ist, wenn es ähnlich zu schon bekannten Motiven ist, die zu TFs mit der gleichen DBD assoziiert sind. Basierend auf diesem Ansatz haben wir eine Statistik entwickelt: Wir haben eine DBD Sammlung zusammengestellt, die schon bekannte TF-Motive Paare enthält und diese nach DBDs sortiert. Mit dieser DBD Sammlung kann man für einen TF ohne vorhergesagtes Motiv, jedoch mit bekannter DBD nachschlagen, welche für Motive zu TFs mit der gleichen DBD assoziiert sind. Mit Hilfe von dieser In-

formation haben wir ein Ähnlichkeitsmaß entwickelt (*domain score*), das zwischen dem zu untersuchend TF und einem Kandidatenmotive berechnet, wie gut das Motive zu den Eigenschaften des TFs passt, die durch die DBD festgelegt wird. Der *domain score* kann als Filter eingesetzt werden, um unwahrscheinliche Motive auszuschließen, oder er wird in einer Metaanalyse genutzt, bei der der p-Wert des *domain score* mit den p-Werten von den etablierten Tools kombiniert wird.
Anhand von menschlichen TF ChIP-seq Datensätzen konnten wir zeigen, dass unsere Methode die Vorhersage von zwei existierende TF motif enrichment Tools, *CentriMo* (Bailey and Machanick, 2012) und *PASTAA* (Roider *et al.*, 2009) verbessert, unabhängig davon ob der Domänewert als Filter oder in einer Metaanlyse benutzt wird (siehe Abbildung 3.3).

Regulatorisch Elemente (REMs), wie z.B. Promotoren oder Enhancer enthalten häufig sehr viele TF Bindestellen. Um die Regulation eines Genes zu verstehen ist es entscheidend zu wissen, wo die REMs eines Genes in der genomischen Sequenz lokalisiert sind. Wir haben den EpiRegio Webserver eingeführt, der REMs enthält, die zu Zielgenen assoziiert sind (see Kapitel 4).
Die REMs wurden basierenden auf menschlichen DNase1-Seq und RNA-Seq Daten für verschiedene Zelltypen und Gewebe mit *STITCHIT* (Schmidt *et al.*, 2021) vorhergesagt. Für jedes annotierte Gen hat *STITCHIT* ein lineares Modell gelernt, um REMs in einer festgelegten Region um das Gen zu identifizieren. Um auch REMs zu bestimmen, die weit von ihrem zu regulierenden Gen entfernt liegen, betrachten wir eine Region von $100,000$ Basenpaare stromaufwärts von der Transkriptionsstartstelle des Genes, das gesamte Gen und $100,000$ Basenpaare stromabwärts von der Transkriptionsendstelle des Genes. Dadurch erhielten wir $2,404,861$ REMs assoziiert zu $35,379$ protein-kodierenden und nicht kodierenden Genen.
Unser EpiRegio Webserver enthält diese REMs und ihre zugehörigen Zielgene für 33 Zelltypen und Gewebe des Roadmap Konsortiums (Kundaje *et al.*, 2015) sowie für 13 Zelltypen des Blueprint Konsortiums (Stunnenberg *et al.*, 2016). Mit unserem öffentlich zugänglichen Webserver sind verschiedene Abfragen möglich: Es ist mögliche für Gene die zugehörigen REMs zu erhalten, oder REMs innerhalb eines genomischen Bereichs zu extrahieren. Zusätzlich berechnen wir für jedes REM einen zelltyp-spezifischen Wert, der die Aktivität des REMs im zu analysierenden Zelltyp oder Gewebe im Vergleich zu allen anderen im Webserver enthaltenen Zelltypen beschreibt. Dadurch kann man REMs basierend auf ihrem regulatorischen Einfluss auf einen bestimmten Zelltypen oder Gewebe priorisieren.
Um den Nutzen unseres Webservers zu verdeutlichen, zeigen wir zwei typische Anwendungsfälle. Mit Hilfe des EpiRegio Webservers konnten wir Gene identifizieren, deren Regulation von CRISPR-Cas induzierten Mutationen in nicht-

kodierenden genomischen Regionen betroffen sind (siehe Abschnitt 4.2). Für diese Analyse haben wir einen CRISPR-Cas screen (Wegner *et al.*, 2019) verwendet, der in menschlichen telomerase-immortalisierte retinale pigmentierte Epithelzellen (hTERT-RPE1) durchgeführt wurde. Das Ziel des CRISPR-Cas screens war es, Mutationen zu identifizieren, die zu einer Chemotherapieresistenz führen. Wegner *et al.* (2019) entdeckten 226 gRNAs, die den nichtkodierenden genomischen Bereiche beeinflussen. Mit Hilfe unseres EpiRegio Webserver haben wir für 13 gRNAs 35 assoziierten Zielgene bestimmten können. Wir haben beobachtet, dass die gRNAs mit mehreren REMs überlappen, was dazu führte, dass wir für jede gRNA mehrere assoziierte Gene erhielt. Bei einer ausführlichen Literaturrecherche konnten wir für einige der identifizierten Gene Hinweise finden, dass sie entweder mit dem Wachstum von Krebs oder Chemotherapieresistenz in Verbindung gebracht werden können (siehe Tabelle 4.1). Zusätzlich haben wir eine TF enrichment Analyse mit den identifizierten REMs, die mit den gRNAs überlappen durchgeführt, um TFs zu bestimmen, die besonderes häufig an diese REMs binden können und dadurch auch durch die induzierten Mutationen beeinflusst sein könnten. Um unsere Ergebnisse einordnen zu können, haben wir auch hier eine Literaturrecherche durchgeführt und für einige TFs einen Zusammenhang mit dem Wachstum von Krebs und Chemotherapieresistenz feststellen können (siehe Tabelle 4.2).
In einem zweiten Anwendungsbeispiel haben wir TFs identifiziert, die an der Genexpressionsveränderung von Perizyten mit verringerte circPLOD2 Expression beteiligt sind. Dafür haben wir mit Hilfe unseres EpiRegio Webservers die REMs der differenziell exprimierten Gene bestimmt und eine TF motif enrichment Analyse durchgeführt. In dieser Analyse konnten wir davon profitieren, dass wir die REMs anhand ihrer Aktivität in Perizyten sortieren konnten. Im Labor von Stefanie Dimmeler konnte nachgewiesen werden, dass einer der von uns identifizierten TFs die Genexpressionsveränderungen in circPLOD2-depletierten Perizyten reguliert (siehe Abschnitt 4.3).

Nicht kodierende genetische Variationen innerhalb von REMs können die Genexpression verändern, indem sie TF Bindestellen modifizieren. Dies kann zu Krankheiten oder unterschiedlichsten Phänotypen führen, daher gibt es Studien, die von den resultierenden funktionalen Konsequenzen berichten (Zhang and Lupski, 2015). Es wurden einige Ansätze entwickelt, um solche regulatorischen SNPs (rSNP) und die TFs, deren Bindestellen betroffen sind zu identifizieren. Wir haben jedoch festgestellt, dass nur wenige dieser Methoden eine statistische Signifikanz für ihren differentiellen TF Bindungswert berechnen. Dies ist jedoch notwendig, um zu entscheiden, welche Ergebnisse signifikant von der angenommen Nullhypothese abweicht, dass der SNP die TF Bindestelle nicht beeinflusst. Wird der differentielle TF Bindungswert als p-Wert angegeben ist ein direkter Vergleich zwischen unterschiedlichen TFs

möglich. Wenn die Methoden einen p-Wert berechnen sind sie oft langsam, besonders für große SNP Datensätze wie zum Beispiel bei genomweiten Assoziationsstudien (GWAS) oder umfangreichen eQTL-Studien.
Wir haben einen maximalen differentiellen TF-Bindungswert (D_{max}) für allgemeine TF Modelle, die bewerten wie gut ein TF an eine Sequenz binden kann, entwickelt (siehe Kapitel 5). Für alle möglichen Teilsequenzen, die den SNP überlappen, berechnen wir einen log-ratio der p-Werte der TF Bindungswerte für den Wildtyp und für die mutierte Sequenz. Der resultierende maximale Wert wird als D_{max} definiert. Wir haben die Verteilung von D_{max} für TF Modelle untersucht, die beispielhaft durch PWMs unterschiedlicher Länge repräsentiert werden und festgestellt, dass sie gut durch eine modifizierte Laplace-Verteilung approximiert werden kann (siehe Abbildung 5.1). Mit Hilfe der modifizierten Laplace-Verteilung ist es uns möglich einen p-Wert für D_{max} zu berechnen, der zwischen unterschiedlichen TFs vergleichbar ist.
Wir haben unseren statistischen Ansatz anhand von *in-vitro*- und *in-vivo* Datensätzen ausgewertet, welche die Qualität unserer Methode zeigt, da eine angemessene AUCPR erreicht wird. Unser statistischer Ansatz verbessert die Vorhersage im Vergleich zu einem etablierten Ansatz geringfügig, jedoch ist er deutlich schneller (siehe Abbildung 5.3).
Wir haben unseren statistischen Ansatz mit zelltyp-spezifischen epigenetischen Daten und zusätzlichen Funktionen in einen Workflow namens Sneep kombiniert. Sneep kann bei der Beantwortung verschiedener Fragestellungen behilflich sein, zwei Anwendungsbeispiele werden im Folgenden kurz erläutert. Zum Einen kann man mit Hilfe unseres Sneep Workflows TFs identifizieren, deren Bindungsstellen häufiger von den zu analysierenden SNPs betroffen sind als erwartet. Dafür haben wir für zwei eQTL Datensätze unseren Sneep Workflow angewendet. Für einige der resultierenden TFs konnten wir in der Literatur Belege finden, dass sie zu den analysierten Zelltypen assoziiert sind (siehe Abschnitt 6.2 und Abbildung 6.2A). Des Weiteren sind die identifizierten TFs signifikant höher exprimiert als TFs, deren Bindungsstellen nicht häufiger als erwartet von den analysierten SNPs betroffen sind (siehe Abbildung 6.2B).
In einem weiteren Beispiel haben wir mit Hilfe von rSNPs und REM-Gen Interaktionen Gene identifiziert, die zu kardiovaskuläre Krankheiten assoziiert sind. Hierfür haben wir unseren Sneep Workflow angewendet um für GWAS von kardiovaskulären Krankheiten rSNPs zu bestimmen, welche wir dann mit Hilfe von bekannten REM-Gen Interaktionen zu Genen in Verbindung bringen konnten. Unter den so ermittelten Genen befindet sich auch das nicht kodierende Gen *IGBP1P1*, für welches im Labor gezeigt werden konnte, das es in einem menschlichen kardiovaskulären Organoid Modell kardiovaskuläre Funktionen verbessert.