COMPUTATIONAL APPROACHES FOR THE PREDICTION OF GENE REGULATORY ELEMENTS AND THE ANALYSIS OF THEIR EVOLUTIONARY CONSERVATION

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

TOBIAS ZEHNDER

Berlin, September 2020

Erstgutachter:	Prof. Dr. Martin Vingron
Zweitgutachter:	Prof. Dr. Boris Lenhard
Tag der Disputation:	21. December 2020

PREFACE

The content of this thesis is the product of the work of the last four years with the objective of obtaining a doctoral degree in science from the faculty of mathematics and computer science at the Freie Universität Berlin. In 2016 I started my PhD in the research group of Martin Vingron under his and Philipp Benner's supervision, together with whom we conceptualized my first project about enhancer prediction. We published the resulting method in *BMC Bioinformatics* early 2019 (Zehnder, Benner, and Vingron, 2019 [1]). In spring 2019 I visited the lab of Boris Lenhard at Imperial College in London for a three months research stay where I learned a lot about conserved noncoding elements, comparative genomics and zebrafish as a model organism. During this time, a fruitful collaboration emerged together with Boris and his post-doc Damir Baranasic in which we developed a method for comparing unalignable genomic sequences between species with large evolutionary distances. This work contributes to a large project with collaborators from different institutions and a manuscript is currently in preparation. Together, the "Berlin" and the "London" projects constitute the main content of my work presented in **Chapters 4** and **5**.

I wrote this thesis in the active voice as in my personal opinion this style of writing is more engaging and allows to deliver a most concise yet intelligible text to the reader. As this document is a single author doctoral thesis and for reasons of consistency, I use the personal pronoun 'I' instead of 'We' whenever not stated otherwise. However, it should be noted and I strongly acknowledge that this work would not have been possible without the support of my supervisors, collaborators and others.

ACKNOWLEDGMENTS

I would like to thank the following people who were pivotal for me and my professional development during the last four years:
Martin Vingron for giving me the opportunity to pursue a PhD in his research group. Martin has provided ideal conditions for my work and sparked my interest in studying enhancers. A special thank goes out to Philipp Benner for his dedicated supervision. Philipp had an open ear for anything anytime and endless patience for explaining. Thank you! Boris Lenhard for hosting me in his research group and for introducing me to the fascinating world of conserved non-coding elements. Kirsten Kelleher for being a great PhD coordinator, proofreading, organizing inspiring events and being an innovator at the institute. Everybody in the IT department of the institute, likely the best in the world. Gal Barel and Roman Schulte-Sasse, the thesis writing team. It was great to have 'companions in misfortune' for mutual motivation during the writing process and inspiration on every level, for example as a first outlet for frustration about the pathological design of R or for discussing elaborate economic models regarding the appropriate pricing of goat cheese salad in the french bakery. Everybody in the Lenhard and Warnecke groups at Imperial College in London for making me feel welcome from day one and including me in all activities. My office mates Lam-Ha Ly, Edgar Steiger, Meng Zhang and especially Robert Schöpflin with whom I was sharing the office during pre-pandemic times and had many inspiring discussions and a shared interest in gene regulation. The WIG team, mainly Anna Ramisch for collaboration and proofreading.
I am greatly thankful to the many people who helped me advance personally and supported me in the past four years:
My brother Sämy for endless discussions about algorithms, life and the universe. My sister Nicole for sharing the adventure of moving to Berlin and her persistent ambition of pretending to understand what I do while truly believing that I dissect viruses. My partner Eva for making me see the positive in everything. Judith for her help with illustrations and design and for her great personal support. My good friends Gal, Roman, Sandy, Dermot and Mike for supporting and motivating me whenever I struggled with work. A big shout-out goes to all my good friends back home in Switzerland who I miss and who came to visit or were hosting me during my visits. My parents for everything. I am lucky and deeply grateful for the privilege of having been raised and supported by you.

TABLE OF CONTENTS

1 INTRODUCTION

The field of molecular biology aims at understanding the processes of the regulation of gene expression in single- and multicellular forms of life as they ultimately shape the identity of cells, tissues, and organisms. Gene expression during embryonic development is a complex and highly orchestrated concert of regulatory processes that culminates in the birth of a new life. Deviations from the regular plan of gene regulation is the breeding ground for both disease and evolutionary change.

Regulation happens at every level during the process of gene expression, and one of the key steps is the transcription of **DNA** to **RNA**. Much of the transcriptional regulation happens not at the gene itself, but at cis-regulatory elements such as promoters, enhancers and silencers. These elements are involved in recruiting the transcriptional machinery to the **transcription start site** of a gene and initiate transcription. As the genetic information in every cell of an individual organism is identical, differential gene activity across tissues is accomplished through epigenetic heterogeneity, i.e. tissue-specific features beyond the **DNA** sequence.

The scope of this thesis is the development of computational methods that describe the epigenomic properties of enhancers and facilitate their identification. Moreover, this work explores the ambiguity of what it means for a genomic element to be evolutionarily *conserved* between species and challenges the current notion that mainly focuses on sequence conservation.

1.1 THESIS OUTLINE

This thesis comprises five Chapters including the current introduction. In **Chapter 2** I will review the general biological background necessary for understanding the biological aspects of this thesis. In addition, I will explain experimental techniques producing the relevant biological data for this dissertation. In **Chapter 3** I will describe the fundamental mathematical concepts on which the subsequently discussed computational methods are based. In **Chapter 4** I will present a computational method for the identification of cis-regulatory elements on epigenomic data using supervised and constricted **hidden Markov models**. I will compare the method to the state-of-the-art, describe its advantages and disadvantages and discuss the properties of the genome-wide predicted elements. In **Chapter 5** I will turn to the subject of evolutionary conservation of regulatory elements. I will present two methods that map non-alignable sequences between two species with large evolutionary distances un-

der slightly different aspects and aim at identifying functionally conserved elements beyond sequence conservation. Finally, I will discuss the work presented in the preceding Chapters and conclude the dissertation with an outlook to potential future directions.

BIOLOGICAL BACKGROUND 2

"The molecules of life are not yet life itself any more than a pile of bricks and lumber is a mansion. At a minimum, life needs a metabolism... [and] the ability to make more of itself - to replicate."
— A. Wagner, Arrival Of The Fittest, 2014.

The field of molecular biology studies - as its name implies - life at the molecular level. Molecular biologists are interested in the structure and composition of cellular components as well as their interactions in diverse processes that together contribute toward an organism's maintenance and proliferation. Proteins are one of the key actors of cellular processes and take on many roles. Structural proteins establish the shape of cells or exert mechanical forces as motor proteins, enzymes are a class of proteins that catalyze chemical reactions, and signaling proteins such as receptors or hormones establish signal transduction mechanisms. Proteins are large macromolecules consisting of amino acid chains, and the particular linear sequence as well as three-dimensional fold that such a sequence entails, gives rise to the protein's function.

The information on how to assemble a protein, i.e. which amino acids to string together in which order, is stored in underlying blueprints, the genes. Genes are functional units on a large linear molecule called **desoxyribonucleic acid (DNA)**. **DNA** consists of two strands of polynucleotides that coil around each other to form a double helix. The nucleotides of each strand are composed of a phosphate-sugar backbone that is bound to that of neighboring nucleotides, and a nucleobase that interacts with its complement on the opposite strand to form a base pair. The four bases in **DNA** are **adenine (A)**, **cytosine (C)**, **guanine (G)** and **thymine (T)**, and base pairing happens between **A** and **T** as well as between **G** and **C** via hydrogen bonds. In a process called replication, the protein complex **DNA** polymerase duplicates the cell's **DNA** before the cell divides into two daughter cells, each containing one copy. That way, genetic information is passed onto the next generation during cell proliferation.

DNA is not only replicated, it can also be transcribed to **ribonucleic acid (RNA)**, which itself can then be translated to synthesize proteins. **RNA** is a molecule similar to **DNA**, but its sugar backbone contains an additional hydroxyl group, hence the name ribose instead of deoxyribose. Moreover, it uses an unmethylated form of **thymine** called **uracil (U)** and often occurs as a single-stranded molecule that folds onto itself. Transcription and translation are the core constituents of a process called gene expression. Together, replication and gene expression constitute a process so central that it has been termed the *central dogma of molecular biology*, a hypothesis pos-

tulated by Francis Crick in 1958 describing the flow of genetic information within a biological system [2]. The basic principle of the central dogma of molecular biology is depicted in **Figure 2.1**.

Today, gene expression is understood to be a complex and highly regulated concert of processes sustaining a cell's functions and thus ultimately essential for the maintenance of life.

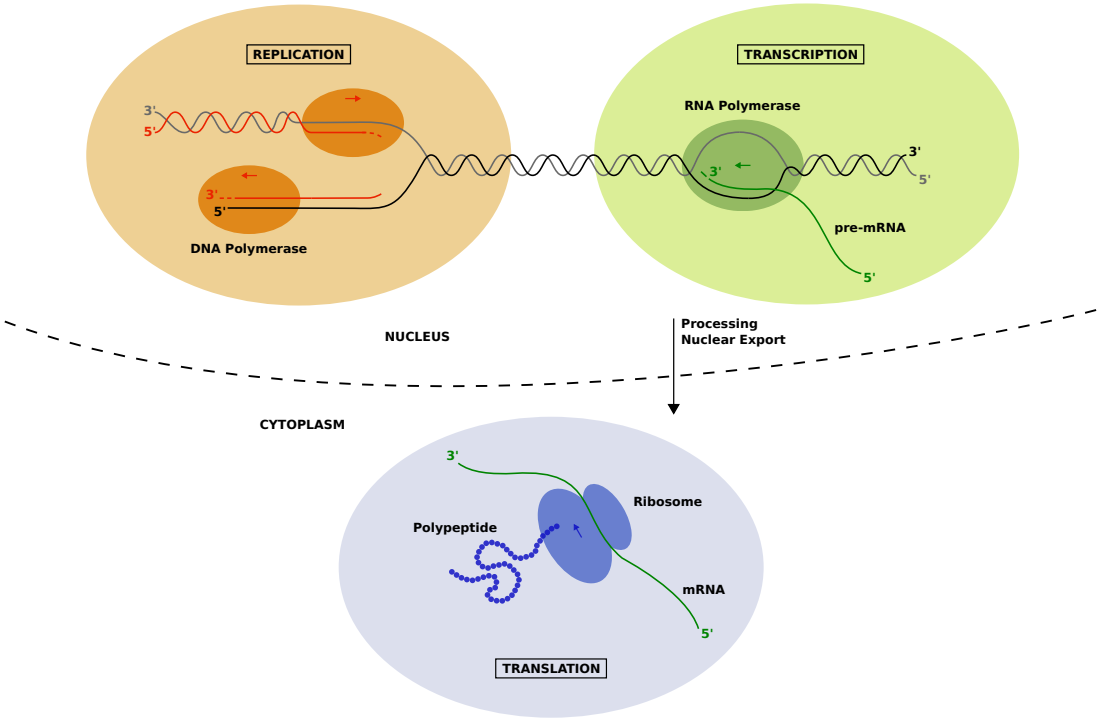

Figure 2.1: The central dogma of molecular biology. The **DNA** double-helix is unwound for replication or transcription which is conducted by different types of polymerases. The transcriptional polymerase produces pre-**messenger RNA (mRNA)**, which is transported from the nucleus to the cytoplasm after processing. There, mature **mRNA** gets translated by ribosomes to polypeptides, i.e. proteins.

2.1 FROM DNA TO PROTEIN - THE CENTRAL DOGMA OF MOLECULAR BIOLOGY

The **DNA** of a human cell - its genome - comprises around 3.1 billion base pairs [3], and only around 1.5% thereof is considered to code for proteins in a total of approximately 20,000 genes [4–7]. The **DNA** of an active gene is read by the protein complex **RNA Polymerase II (RNAP II)** and used as a template to create **RNA** in a process called transcription. **RNA** comes in many different types: **mRNA** is the transcript of protein-coding genes and is transported from the nucleus to the cytoplasm where ribosomes translate it into protein according to the genetic code. However, the majority

of transcripts does not code for proteins [8]. The ribosome itself is a big complex composed of both ribosomal proteins and **ribosomal RNAs (rRNAs)**, and the process of translation involves the transfer of amino acids carried by **transfer RNAs (tRNAs)** to the nascent peptide chain. Other types of **RNA** are involved in the regulation of transcription (e.g. **long non-coding RNAs (lncRNAs)**) as well as the regulation of gene expression at the post-transcriptional level (e.g. **microRNAs (miRNAs)** and **small interfering RNAs (siRNAs)**).

Gene expression is regulated on all levels, i.e. before and after transcription as well as on a post-translational level. This thesis focuses on the transcriptional aspect of gene regulation. In the remaining part of the Section I will describe the processes of transcription in more detail and introduce the concept of transcriptional regulation.

2.1.1 TRANSCRIPTION

Transcription can be divided into three major steps: initiation, elongation and termination. Transcription initiation happens at gene promoters, short cis-regulatory elements located in proximity to the **transcription start site (TSS)** of the gene they regulate. In this context, *cis* means acting on the same **DNA** molecule, i.e. the same chromosome. In contrast, trans-acting factors, e.g. **transcription factors (TFs)**, interact with different molecules. Promoters harbor binding sites for **general transcription factor (GTF)**, which perform a series of actions: First, they recruit **RNAP II** to the promoter, forming and stabilizing the initiation complex. The initiation complex then unwinds the **DNA** forming a transcription bubble in preparation for gene transcription.

Eventually, the positive transcription elongation factor P-TEFb phosphorylates the C-terminal domain of **RNAP II**, causing it to undergo a series of conformational changes that release it from the promoter to begin transcription elongation, at which point the **GTFs** are released from the promoter and available for a new round of transcription initiation. During transcription elongation, **RNAP II** slides along the gene body and transcribes the **DNA** template strand to **mRNA**. In humans, transcription elongation happens at a speed of around 3.3 - 3.8 **kilo base pairs (kbp)**/min [9]. The average human gene is 28 **kbp** long [5], and its transcription therefore takes around 7-8 minutes. The largest human gene DMD covers 2.3 **mega base pairs (Mbp)** and it takes more than 10 hours to transcribe it. However, transcription can be initiated before the previous transcription terminated and thus often takes place in parallel.

During transcription termination, the nascent transcript is cleaved and prepared for downstream processing. For example, **mRNAs** are subject to intron splicing, 3'-polyadenylation and 5'-capping before preparing them for nuclear export and translation in the cytoplasm. Non-coding transcripts are frequently processed during or after transcription termination too, e.g. different types of **rRNAs** are the product of cleaving and chemically modifying a large precursor **rRNA**.

2.1.2 TRANSCRIPTIONAL REGULATION

In eukaryotes, transcription is regulated during all major steps. The mechanisms are manifold and involve trans-acting factors such as **TFs** as well as cis-acting regulators and chromatin structure.

TRANSCRIPTION FACTORS **TFs** belong to the class of **DNA**-binding proteins and regulate gene expression. They are highly diverse in terms of structure and **DNA** binding mechanisms and group into different families that are defined based on their **DNA-binding domain (DBD)** such as helix-turn-helix, zinc finger, homeodomains, leucine zipper, helix-loop-helix and β-sheet proteins (see [10] for review). Weirauch and Hughes [11] catalog 91 different types of **DBDs** across different domains of life, with binding preference, biological role and regulatory targets still unknown for the majority of eukaryotic **TFs**. In human, there are more than 2600 proteins with **DBDs** [12], suggesting a high complexity of gene regulation through **TFs**. The basic function of a **TF** is the recognition of **DNA** sequence patterns that allows it to bind at specific genomic locations. Subsequently, it either promotes or represses the recruitment of **RNAP II** to the target gene promoter, in many cases in concert with other **TFs** and through coactivators, chromatin remodelers and **histone modification (HM)** enzymes.

CHROMATIN STRUCTURE Eukaryotic **DNA** is wrapped around protein complexes consisting of eight histones in order to compact the **DNA** enough to fit in the nucleus. Histones are the protein components of chromatin and have an arginine- and lysine-rich N-terminus, conferring them a basic and positively charged character that allows a tight interaction with the negatively charged **DNA**. Together, histones and **DNA** form the lowest subunit of chromatin, the nucleosome. The presence of nucleosomes has an effect on the transcription dynamics. For example, nucleosomes slow down transcription elongation as they impede the progression of **RNAP II**. Moreover, highly compacted **DNA** is not accessible for most **TFs**, and it requires the action of

so-called pioneer factors and chromatin remodeling complexes to open up chromatin in order to present the **transcription factor binding site (TFBS)** to other **TFs**. Chromatin state and structure are therefore crucial elements in the regulation of gene transcription.

There are several ways to alter chromatin state. A core nucleosome consists of two copies each of the four different histones H2A, H2B, H3 and H4. The two latter have long N-terminal tails, and their residues are subject to covalent modifications by histone-modifying enzymes. For example, **lysine 27 at histone 3 (H3K27)** can be acetylated or methylated, altering the characteristics of the chromatin at that location. At that residue, acetylation neutralizes the positive charge of lysine and thus weakens the interaction between the negatively charged **DNA** and the histone [13]. Methylation has the opposite effect, facilitating chromatin compaction. The addition of linker histone H1 to chromatin results in the nucleosomes being positioned in regular arrays, forming tightly compacted chromatin fibers that are not accessible for the transcription machinery. The regulation of chromatin accessibility through nucleosome positioning and modification thus composes a powerful way to regulate gene expression at the transcriptional level.

Chromatin structure affects transcription not only at the level of nucleosomes. **DNA** is a linear molecule, but chromosomes are compartmentalized at multiple spatial scales [14]. This arrangement in the nucleus results in the spatial proximity of regions that are far apart on the linear **DNA** sequence. Not surprisingly, this spatial organization of chromatin is not random. In 2012, it was discovered that eukaryotic genomes form chromosome territories, within which **DNA** sequences interact more frequently with each other than with sequences from outside such domains, termed **topologically associated domains (TADs)** [15, 16]. Chromatin that interacts with the nuclear lamina forms **lamina associated domains (LADs)** and becomes transcriptionally silent. For a review on **TADs** and **LADs**, see [17]. The full range of functions of **TADs** is still heavily studied, but it has been observed that interactions of promoters with another type of cis-regulatory elements, so-called enhancers, does not happen across **TAD** boundaries. Because enhancers are at the very center of this thesis, I will introduce them separately in the subsequent Section.

2.2 ENHANCERS

Enhancers are short sequences of **DNA** involved in the transcriptional regulation of gene expression. In contrast to promoters, their position is not constrained to a linear proximity to a gene's **TSS**, and they do not necessarily regulate the closest gene. For

example, in mouse, the gene *Shh* and its ZRS enhancer are located approximately 1 **Mbp** apart [18]. However, enhancers are thought to interact with a gene's promoter by coming into close spatial proximity through the three-dimensional chromatin architecture.

The discovery of enhancers dates back to 1981, when Banerji, Rusconi, and Schaffner [19] first described the "enhancement of globin gene expression" in transfected HeLa cells. Transcription of the cloned rabbit hemoglobin β1 gene was elevated 200-fold when the gene-plasmid recombinants also contained the viral DNA segment SV40. The authors found that the enhancing activity of SV40 was independent of orientation or exact position relative to the **TSS** of the target gene, and concluded that the activation of genes by this class of **DNA** elements, transcriptional enhancers, ought to be a general mechanism for gene regulation. These findings have fueled research of transcriptional regulation in general and enhancers in particular.

There have been many attempts to describe the properties of enhancers since they were discovered in the early 1980s. To date, there is still no well-defined description of an enhancer that would make their identification unambiguous. However, there have been myriads of studies reporting different enhancer features. In the remaining parts of this Section I will discuss the properties of enhancers, their mechanistic function in targeting a specific gene promoter, how they tightly regulate gene transcription in a cell type-specific manner during development and differentiation, their evolution, as well as their role in disease.

2.2.1 ENHANCER FUNCTION

Despite the lack of feature descriptions encompassing all enhancers, there is agreement on the definition of their function: an enhancer elevates the rate of transcription initiation of its target gene, and it does that by recruiting the essential components of the transcription machinery [20]. Enhancers have been found to target multiple genes depending on the cell type, and similarly, genes can be regulated by multiple enhancers, either at the same time, or separately in a cell type-specific manner [21]. Walter Schaffner, the head of the team that discovered the β-globin enhancer in 1981, wrote in a 2015 review that an enhancer is a "**DNA** platform that interacts with a multitude of transcriptional regulatory proteins" [22]. These proteins, so-called activators, either directly recruit **GTFs** and **RNAP II** to the promoter, or they do so via coactivators. One of the most prominent coactivators is Mediator, a large protein complex with a variable subunit composition involved in multiple steps of **RNAP II** regulation [23]. With a molecular weight of about 1.4 **megadalton (MDa)**, Mediator is

roughly three times the size of **RNAP II** [24, 25]. During transcription initiation, Mediator, **GTFs** and **RNAP II** form protein-protein interactions. The exact mechanism by which activators and coactivators regulate transcription initiation remains poorly understood, however, different models have been proposed.

According to the *recruitment* model, enhancers are simple binding platforms for **TFs** that establish the assembly of the transcriptional machinery [20]. An enhancer's activity depends on many parameters, e.g. the quantity and quality of **TF** binding events, the simultaneous presence and interaction of multiple **TFs**, their arrangement and orientation as well as structural properties of the subjacent **DNA** itself. As a basic principle, enhancers decipher a combinatorial code, a function of the described parameters [26–29]. Multiple suggestions regarding an enhancer's integrative function of those parameters have been proposed.

The *enhanceosome* model, first described by Bazett-Jones et al. [30] in 1994, suggests the requirement of strict **TF** positioning in a specific order on the enhancer. This facilitates protein-protein interactions between the **TFs**, leading to the formation of a higher-order protein complex that recruits **RNAP II**. The necessity of all *enhanceosome* components to be present at the same time in order to confer **TF** cooperativity provides a sharp, binary transcriptional switch for the activation of gene expression [31]. The *enhanceosome* has been described in detail on the transcriptional activation of the IFN-β gene [32–34].

Most developmental enhancers, however, do not exhibit the highly ordered **TFBSs** typical for the *enhanceosome* model, but rather allow for the additive and even independent contribution of **TFs** to gene activation [31, 35]. This led to the description of the *billboard* model which allows for a flexible positioning of **TFBSs** [35]. According to this model, enhancers provide information displays, i.e. billboards, which are interpreted by the basal transcription machinery. More than only through additivity, it seems possible that in some cases indirect **TF** cooperativity can be achieved without the need to constrain their relative positioning.

Moreover, some enhancers are enriched for particular **TFs** without providing the appropriate binding sites, but instead through protein-protein interactions with other **TFs** for which binding sites are present. This can be viewed as a model in which a subset of the required **TFs** act as coactivators, i.e. they require the **DNA**-binding property of other **TFs** in order to be directed to the site of transcription activation, known as the ***TF** collective* model [36].

Very recently, there has been a study suggesting a *soft syntax* model of regulatory integration at enhancers, which can be understood as an intermediate of the *billboard* and the *enhanceosome*, agreeing with the *recruitment* and *collective* models [37]. For example, one of the rules of *soft syntax* appears to be an ordered spacing of binding motifs with helical periodicity, i.e. distances of ~10.5 **base pair (bp)**. These findings are in line with the *collaborative nucleosome competition* model that states that **TF** binding to nucleosomal **DNA** is inherently cooperative due to the **TFs**' competition with nucleosomes [38, 39], representing yet another mechanism of **TF** cooperativity that does not require protein-protein interactions.

For a review on enhancer architecture and **TF** integration models, see [40]. These recent findings are consistent with the previously reported observation that **TFs** can not only recognize **DNA** sequence, but also **DNA** shape, and that this shape readout can reach beyond the core **TFBS** [41]. **DNA** shape is highly dependent on both nucleotide composition and chromatin state, i.e. how the **DNA** is organized into the three-dimensional volume of the nucleus.

2.2.2 ENHANCER CHROMATIN

Some of the coactivators function as chromatin remodelers and histone modification enzymes. They enable relaxation of the chromatin from its tight packaging around nucleosomes and even allow histones to be removed, making the **DNA** at both enhancers and promoters accessible for the binding of other factors, and ultimately available for unwinding the double strand in order to form the transcription bubble. Chromatin accessibility can be measured using experimental assays such as **DNase-seq** [42, 43] or **ATAC-seq** [44], and I will address them in **Subsection 2.4.2**.

Histone modifications (**HMs**) do not only alter the physical configuration of chromatin, they can also be recognized and read by other factors, which led to the proposition of an epigenetic code, the histone code [45, 46]. The hypothesis suggests that, although the **DNA** remains the same across different conditions, developmental stages, cell types and tissues, the epigenetic code could confer specificity. Consistent with that, Heintzman et al. [47] reported in 2009 that **HM** patterns are highly cell type-specific at enhancers, and less so at promoters, suggesting that it is the enhancers that confer specificity to gene transcription.

As of today, there is a plethora of **HMs** known to correlate with particular genomic functions. For example, **H3K36me3** is a mark broadly spread over actively transcribed gene bodies and is linked to transcriptional elongation by **RNAP II** [48]. As briefly

touched upon in **Subsection 2.1.2**, **acetylation of histone 3, lysine 27 (H3K27ac)** generally relaxes chromatin and is associated with transcriptionally active regions, whereas trimethylation of the same residue (**H3K27me3**) has the opposite effect. Correspondingly, both active enhancers and promoters are associated with high levels of **H3K27ac** [49, 50]. Moreover, the degree of methylation at **H3K4** has been understood to allow the distinction of the two cis-regulatory elements. Active promoters often have the residue trimethylated (**H3K4me3**) [51, 52], whereas at enhancers it is typically monomethylated (**H3K4me1**) [53, 54]. Consistent with that, methylated **H3K4** has been shown to facilitate subsequent chromatin remodeling such as histone acetylation by the **histone acetyltransferase (HAT)** p300 [55–57]. Enhancers with the active mark **H3K4me1** but with their **H3K27** trimethylated instead of acetylated were found to be in an intermediate state between active and repressed, termed *poised* [58, 59]. These enhancers are silent, but ready to be activated quickly by acetylation of **H3K27**.

As a consequence of both the considerable advances in high-throughput techniques such as **ChIP-seq** for the genome-wide mapping of *in vivo* protein-**DNA** interactions (see **Subsection 2.4.3**) as well as increasingly available antibody reagents, the list of studied **HMs** has become very long. Zhao and Garcia [60] provide a catalog of the vast number of studied **HMs**.

2.2.3 DNA METHYLATION

Not only the protein components of chromatin but also the **DNA** sequence itself can be subject to post-translational modifications. Methylation of cytosine, one of the four bases of **DNA**, is a widespread epigenetic modification across all domains of life. Methylated **cytosine** is less stable than in its unmethylated form and prone to deamination, resulting in a mutation to **thymine**. In mammals, **DNA** methylation is predominantly present at CpG dinucleotides and is thought to alter the underlying sequence's activity without changing the sequence itself [61, 62]. As a consequence, CpG-dinucleotides are about four times less frequent than expected in the human genome except for distinct loci termed CpG-islands, which often co-occur with gene promoters [63]. In fact, the association of demethylation of CpG-dinucleotides at promoters with increased transcriptional activity has allowed many promoters to retain expected CpG frequencies and endowed them with an additional layer of regulation.

Enhancers typically also exhibit reduced **DNA** methylation, albeit not to the same extent as promoters. Stadler et al. [64] reported the existence of **low-methylated regions (LMRs)** with an average methylation level of 30%, representing CpG-poor distal reg-

ulatory regions exhibiting enhancer-typical chromatin marks. Moreover, enhancers co-localize with **differentially methylated regions (DMRs)**, i.e. regions with differential patterns of **DNA** methylation across tissues and developmental stages [62]. Cell type-specific hypomethylation patterns have been reported to correlate with **H3K27ac** and reflect cell type-specific enhancers [65]. **DNA** methylation patterns are often disturbed in cancer epigenomes, leading to uncontrolled transcription of oncogenes or the silencing of tumor suppressor genes [66, 67]. For a review on **DNA** methylation, see [68].

2.2.4 ENHANCER TRANSCRIPTION

CBP and p300 are two proteins from the same family of coactivators. They both act as a **HAT** and deposit **H3K27ac** mainly at enhancers [49, 50, 53, 69]. CBP has been shown to recruit **RNAP II** to the site of transcription initiation at the promoter. As a result of the close spatial proximity of enhancer and promoter during that process, it should not surprise that **RNAP II** occasionally transcribes the enhancer instead of the promoter. Kim et al. [70] found that in mouse cortical neurons, about 25% of enhancers are transcribed bi-directionally, yielding short non-coding transcripts. In Drosophila, over 94% of enhancers in regions of accessible chromatin have been found to be transcribed with a minimum of five reads, indicating that weak enhancer transcription might be the rule rather than the exception, but that some of it might be missed by experimental assays lacking sensitivity. Initiation architecture as well as frequencies of core promoter elements happen to be highly similar between enhancers and promoters [71], however, the two elements diverge in post-initiation transcript stability [71]. Enhancer transcription happens both uni- and bidirectionally [72, 73], and it is unclear whether the transcripts [74, 75] or the process of transcription itself [76] are functionally important for the process of gene regulation, or if enhancer transcription is merely a consequence of the spatial proximity of **RNAP II** to the enhancer's accessible chromatin [77]. The answer is likely a combination of all of these hypotheses, and not all enhancers are likely to be governed by the same processes.

Despite recent reports that enhancer transcription is less predictive of enhancer activity than e.g. the presence of particular patterns of **HMs** [78], the existence of a link between an enhancer's transcription and its activity is widely accepted. Several experimental techniques measure the production of **enhancer RNAs (eRNAs)**, e.g. **cap analysis gene expression (CAGE)** [79], PRO-seq [80], GRO-seq [81] and Start-seq [82, 83]. In particular, **CAGE** was used by the **FANTOM** consortium to establish a comprehensive atlas of transcribed enhancers in multiple organisms and cell types [84]. I will describe the technique in more detail in **Subsection 2.4.4**.

2.2.5 ENHANCER CONSERVATION

Genomic portions that confer function essential to the viability and fitness of an organism tend to be evolutionarily conserved [85, 86]. Mutations of such **DNA** sequences may lead to a reduced fitness and are thus naturally selected against, a phenomenon known as negative or purifying selection [85]. Conversely, evolutionary conservation of a **DNA** sequence can indicate its functional importance [86], although it does not necessarily prove it [87, 88].

The different models of how an enhancer integrates the regulatory input from its binding **TFs** implicate different evolutionary constraints on enhancer sequence. Enhancers acting according to the *enhanceosome* model rely on a strict order of **TFBSs** and are thus expected to be conserved across species. This is in fact the case for developmental enhancers, which are often deeply conserved across large evolutionary distances and distribute non-randomly in clusters around the genes they regulate. Such regions of unexpected high sequence conservation and their clusters have been termed **conserved non-coding elements (CNEs)** [89–91] and **genomic regulatory blocks (GRBs)** [92], respectively, and I will address them separately in the next Section.

Enhancers under the *billboard* and *collective* models possess much more plasticity and are expected to be less conserved in sequence. Of course, enhancers might also regulate species-specific traits and exhibit no sequence conservation as they might have evolved only recently [93]. Moreover, some enhancers act alone and are essential while others are dispensable as their loss might be buffered by redundant enhancers [94].

Enhancers with low sequence conservation can mean two things: either the enhancer is indeed specific to a particular species or clade, or the enhancer's function is conserved, but that conservation is not reflected in the sequence, consistent with the more plastic *billboard* and *collective* models. Moreover, as opposed to promoters, which need to be located near the **TSS** of a gene, enhancers are not constrained in their linear genomic position as long as they manage to spatially fold to their target promoter. Under these circumstances, some enhancers possess more freedom to mutate while conserving ancestral function. One proposed mechanism is compensatory **TFBS** turnover as has been described in detail for the even-skipped stripe 2 enhancer (S2) in various Drosophila species [95]. Another is the actual movement, or in evolutionary terms the turnover of the actual position of an enhancer. This has been described for the Drosophila gene *yellow* regulated by an enhancer which is found at variable genomic positions in different Drosophila species [96]. Taken together, enhancers are a very

diverse type of genomic elements with various modes of action and therefore display heterogeneous levels of sequence conservation [97], with some exhibiting functional rather than sequence conservation [86, 98–101].

2.3 SHARED SYNTENY OF CONSERVED NON-CODING ELEMENTS IN GENOMIC REGULATORY BLOCKS

In 1985, Yaffe et al. [102] investigated sequence conservation in **mRNA** coding for homologous actins and observed high degrees of homology in parts of the 3′ **untranslated region (UTR)**, i.e. the part of the transcript downstream of the translation stop codon. A few years later, Lemaire, Heilig, and Mandel [103] made a similar observation in the 3′ **UTR** of the dystrophin gene, followed by other studies reporting the existence of high sequence conservation in introns, 5′ and 3′ **UTRs** and other non-coding regions [104–107]. The development of **next generation sequencing (NGS)** technologies greatly facilitated comparative genomics studies and led to three studies in quick succession that systematically analyzed and described so-called **conserved non-coding elements** (**CNEs**) genome-wide [89–91].

CNEs are genomic stretches of **DNA** exhibiting striking levels of sequence conservation that exceed those of protein coding genes with some of them being at least partially conserved over hundreds of millions of years [90]. The conservation levels of **CNEs** are unexpected under neutral selection, and it has been reported that they are subject to negative selection that is much stronger than that in protein coding genes due to the degeneration of the genetic code [108]. To date, a conclusive model explaining this degree of sequence conservation in **CNEs** remains elusive. Graphic examples of the degree of sequence conservation across multiple species of an example gene exon and an example **CNE** are given in **Figure 2.2 A** and **B**, respectively.

CNEs are not distributed randomly in the genome but tend to cluster into highly syntenic structures, i.e. they co-localize in both species with a conserved order, hence their genomic positions are collinear [109]. Moreover, such arrays of **CNEs** are often found around key developmental genes, which they are expected to regulate as enhancers [90]. Therefore, these clusters were termed **genomic regulatory blocks** (**GRBs**) [92]. Besides the target gene regulated by the **CNEs**, **GRBs** often also contain bystander genes, i.e. genes whose syntenic positions were maintained solely due to their interspersion with **CNEs**, but whose expression and function are distinct from those of the target gene due to a different promoter architecture [110, 111]. A visual interpretation of the **GRB** model is given in **Figure 2.2 C**.

GRBs have been shown to coincide with **TADs**, suggesting a role of **CNEs** in the three-dimensional organization of chromatin [112]. Conversely, not all **TADs** coincide with **GRBs**. Those who do are typically larger and gene-sparser as well as attributed with higher strengths of intra-**TAD** interactions than those who do not contain clusters of **CNEs**. Also, target genes in **GRB**-overlapping **TADs** tend to exhibit a higher cell-type-specificity. As these two types of **TADs** often alternate on the linear genome, it has been proposed that the insulation of the **TADs** that do not overlap **GRBs** is simply a consequence of their adjacency to strongly interacting **TADs** [112]. For a review about **CNEs** and **GRBs**, see [113, 114].

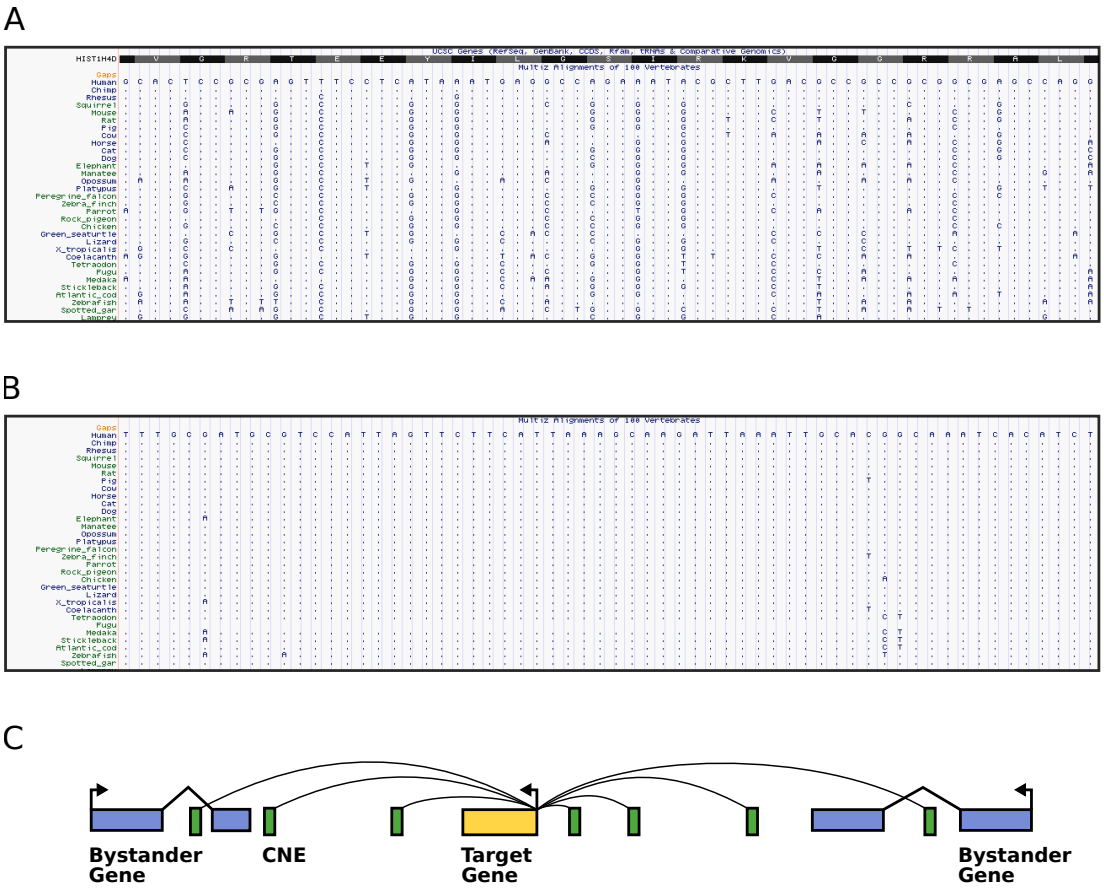

Figure 2.2: CNEs exhibit high degrees of conservation and cluster in **GRBs**. **A** Multiple alignment of an example coding region (exon of HIST1H4D, hg19 at chr6:26,189,130-26,189,194). Dots represent identical nucleotides with respect to the hg19 reference. **B** Multiple alignment of an example **CNE** (hg19, chr3:180,462,367-180,462,428). **C** Schematic illustration of the **GRB** model.

2.4 EXPERIMENTAL TECHNIQUES IN FUNCTIONAL GENOMICS

The field of functional genomics aims to understand an organism's phenotype as a function of its genotype. Different experimental techniques have emerged that allow investigating particular aspects such as gene expression, protein interactions and many more. In this Section I will introduce the essential techniques giving rise to the data that I used in my projects and refer to in other parts of this thesis.

2.4.1 DNA SEQUENCING

As a consequence of the central role of **DNA** in functional genomics, **DNA** sequencing is one of its most central techniques, and many experimental methods incorporate a step of **DNA** sequencing in their protocols. Historically, **DNA** sequencing techniques were developed after the first published sequencing methods for proteins in the 1950s [115] and **RNA** in the 1960s [116]. The first successful attempt to sequence **DNA** dates back to the late 1960s when Wu and Kaiser [117] sequenced the 12 **bps** of the cohesive ends of bacteriophage lambda. In 1973, Gilbert and Maxam [118] published the 24 **bp** long nucleotide sequence of the lac operator. The applied methods were extremely cumbersome and time-consuming, and it was not before 1977 that two methods with drastically improved speed were presented: The chemical cleavage procedure by Maxam and Gilbert [119], and the chain terminator procedure by Sanger, Nicklen, and Coulson [120].

Within the next decade, multiple methods and protocols were developed that refined and automated the existing sequencing techniques [121–123]. By the year 2000, a second generation of sequencing methods emerged, termed **next generation sequencing** (**NGS**) or **high-throughput sequencing (HTS)**. The latter term is owed to the main difference of **HTS** technology to previous methods, namely that it is highly scalable, allowing the sequencing of an entire genome in one go. Today, sequencing by synthesis is one of the most widely used short-read **HTS** methods and universally known as Illumina sequencing [124]. The main steps in the Illumina protocol are as follows: First, **DNA** is fragmented yielding small pieces of a few hundred **bps**. The fragments are clonally amplified in a highly parallel process called bridge amplification. Then, the single **DNA** strands are complemented using fluorescently tagged nucleotides during the sequencing step. The addition of each nucleotide emits a characteristic fluorescent signal that allows deciphering the composition of the **DNA** sequence.

Paired-end sequencing incorporates the sequencing of fragments from both ends and thereby further improves the quality of the sequencing data. Long-read sequencing

techniques such as Nanopore and **single-molecule real-time sequencing (SMRT)** facilitate *de novo* genome assembly at the expense of a decreased accuracy [125, 126]. For a review on the history of **DNA** sequencing, see [127].

2.4.2 CHROMATIN ACCESSIBILITY ASSAYS

Regulatory activity requires the ability of **TFs** to bind and interact with chromatinized **DNA**. This is highly dependent on the degree of chromatin accessibility which is determined by nucleosome occupancy and organization. In 1973, Hewish and Burgoyne [128] described the use of **DNA** endonucleases to fragment chromatin and observed periodic hypersensitivity across the genome. Since then, technological advances have led to today's possibility of measuring chromatin accessibility genome-wide with relatively few requirements in terms of biological material and labor. In the following paragraphs I will describe two of them: **DNase I hypersensitive site sequencing** (**DNase-seq**) and **assay for transposase-accessible chromatin using sequencing** (**ATAC-seq**). For a comprehensive review on chromatin accessibility and designated experimental assays see [129].

DNASE-SEQ **Deoxyribonuclease I (DNase I)** is an enzyme with endonuclease activity which preferentially cleaves nucleosome-depleted **DNA**. It generally does not have a strong sequence preference and thus binds **DNA** relatively nonspecifically, although there have been reports indicating low levels of sequence specificity leading to some bias in resulting data sets [130]. Applying **DNase I** to a population of cells therefore results in short fragments of **DNA** whose ends correspond to the site of digestion, also-called **DNase I hypersensitive site (DHS)**. In 2006, two studies provided the first genome-wide measurements of DNase hypersensitivity [42, 131]. In 2008, Boyle et al. [43] adapted the workflow and presented **DNase I hypersensitive site sequencing (DNase-seq)**, which adds a step of high-throughput sequencing of the fragments. Sequencing and subsequent alignment back to the genome thus allowed quantifying the relative abundance of accessible chromatin throughout the genome. Song and Crawford [132] provide a complete and improved description of the protocol.

ATAC-SEQ **Assay for transposase-accessible chromatin using sequencing (ATAC-seq)** protocols were first presented in 2013 by Buenrostro et al. [44] as an orthogonal approach to **DNase-seq**. They use hyperactive Tn5, a prokaryotic enzyme of the class of transposases, which have been shown to integrate into nucleosome-free regions

in vivo [133]. In **ATAC-seq**, genetically engineered Tn5 transposes into regions of accessible chromatin and ligates high-throughput sequencing adapters to these regions. The presence of nucleosomes sterically hinders such transposition, making it less likely to happen in nucleosome-dense regions. While **DNase-seq** requires millions of cells, **ATAC-seq** can be performed on 500-50,000 cells [44]. It is a simple procedure executed in less than two hours, which stands in high contrast to the multiday protocol of **DNase-seq** [129]. The simplicity, time-efficiency and the fact that **ATAC-seq** data correlates well with **DNase-seq** let it become the state-of-the-art technique for measuring chromatin accessibility [44, 134]. In recent years, modifications of the protocol yielded **single-cell ATAC-seq (scATAC-seq)**, enabling the analysis of accessible chromatin in heterogeneous cell populations at single-cell resolution [135]. For a review on **ATAC-seq**, see [136].

2.4.3 CHIP-SEQ

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an experimental technique used to identify the location of protein-**DNA** interactions on a genome-wide scale and was first described in 2007 [137, 138]. It combines **chromatin immunoprecipitation (ChIP)** and high-throughput sequencing to map binding sites of **TFs**, **HMs** and other **DNA**-binding proteins. **ChIP** is a procedure in which proteins and chromatin are covalently cross-linked at their current site of interaction using formaldehyde. This creates a snapshot of protein-**DNA** interactions in a given cell population. The **DNA** is then fragmented using sonication, followed by isolation of the fragments bound to the protein of interest using antibodies. **DNA** recovery and purification is achieved by the reversal of the protein-**DNA** cross-links. After that, **DNA** is extracted and prepared for high-throughput sequencing. Genome mapping of sequencing reads represent the last step, yielding quantitative maps of genome-wide protein-**DNA** interactions.

It has been reported that highly transcribed and thus accessible chromatin is more selectively recovered during immunoprecipitation, potentially creating biologically meaningless artifacts in **ChIP-seq** data. Other studies show that nearly every step of the protocol is prone to bias [139, 140]. Moreover, low signal-to-noise ratio and a resolution limited to above 100 **bp** sparked the development of improved techniques such as ChIP-exo [141] and CUT&RUN [142]. Regardless, **ChIP-seq** has become the standard method for profiling protein-**DNA** interactions and is one of the most established and widely used techniques in the field. For a review, see [143].

2.4.4 CAGE

In 2003, Shiraki et al. [79] developed **cap analysis gene expression** (**CAGE**), an experimental method to capture the present state of the transcriptome in a cell population of interest. In short, **CAGE** identifies a transcript's 5′ cap, a particularly modified **guanine** nucleotide at the 5′ end of stable **mRNA** transcripts [144]. In 2012, Takahashi et al. [145] described an optimized protocol in which 5′ capped transcripts are reverse-transcribed into **complementary DNA (cDNA)** which is then cleaved by a restriction enzyme 27 nucleotides downstream of the 5′ end. These so-called **CAGE** tags are then amplified using **polymerase chain reaction (PCR)** and subsequently sequenced on the Illumina platform.

Since the introduction of **CAGE**, several improved protocols have been developed that for example no longer require cleavage and **PCR** amplification (nAnTi-CAGE [146]), or operate on a fraction of the input material required in early protocols thanks to a dramatically increased sensitivity (SLIC-CAGE [147]). **CAGE** provides a method for identifying **TSSs**, characterizing promoter usage and determining the initiation sites of both coding and non-coding **RNAs**.

3 MATHEMATICAL PREREQUISITES

3.1 PROBABILISTIC MODELS

Probabilistic models are used to model phenomena and incorporate random variables and probability distributions in order to predict the outcome of an event. In this Section I will give an introduction into probabilistic models relevant for the work presented in this thesis.

3.1.1 STATISTICAL MODELS FOR READ COUNT DATA

High-throughput functional genomics assays followed by sequencing are among the most popular techniques in molecular genetics. They typically produce read count data that can be used to quantify biological phenomena. The distribution of the read count data produced by the final sequencing step is often approximated by a suitable statistical model. Such a model needs to fit the characteristics of the read count data, which is discretely distributed over an unbounded positive range with the sample variance exceeding the sample mean. In other words, read count data is overdispersed with respect to the Poisson distribution. Gierliński et al. [148] describe the suitability of multiple statistical models to approximate the distribution of RNA-seq read count data and find that the negative binomial and the log-normal distributions perform best. I use these distributions for modeling read count data from **ChIP-seq** and **ATAC-seq** experiments as explained in **Section 4.2**. In the following paragraphs I will formally describe the negative binomial and the log-normal distributions.

THE NEGATIVE BINOMIAL DISTRIBUTION The negative binomial distribution is a discrete probability distribution of the number of successful Bernoulli trials x before the occurrence of a number of failures r. With p denoting the probability of success, the probability mass function of a negative binomial distribution is

$$f_{NB}(x; r, p) = \binom{x + r - 1}{x} (1 - p)^r p^x. \tag{3.1}$$

With the mean and variance being $\mu = \frac{pr}{1-p}$ and $\sigma^2 = \frac{pr}{(1-p)^2}$, respectively, we can parameterize the distribution differently and get

$$f_{NB}(x; \mu, r) = \frac{\Gamma(r + x)}{\Gamma(r)x!} \left(\frac{\mu}{\mu + r}\right)^x \left(\frac{r}{\mu + r}\right)^r, \tag{3.2}$$

with $\Gamma(n) = (n - 1)!$ being the gamma function for all natural numbers $n \in \mathbb{N}$.

THE LOG-NORMAL DISTRIBUTION The log-normal distribution is a continuous probability distribution used to describe natural phenomena, e.g. the survival rate of bacterial spores in disinfectants [149], or the blood pressure of adult humans [150]. The logarithm of a log-normally distributed variable follows a normal distribution. The probability density function for the log-normal distribution is

$$f_{LN}(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right). \tag{3.3}$$

3.1.2 MAXIMUM LIKELIHOOD ESTIMATION

The problem of understanding biological phenomena often involves the underlying assumption that a particular phenomenon can be described sufficiently by a statistical model. For example, we might observe some data that we would like to model with a certain type of probability distribution. Once we are set on a model we intend to infer the optimal model parameters that best explain the observed data. **Maximum likelihood estimation (MLE)** is a procedure for that purpose. It maximizes a likelihood function so that under the assumed model the observed data is most probable. **Maximum a posteriori estimation (MAP)** presents an alternative to **MLE** which proves particularly useful when observed data size is small and prior information about the model parameters is available. However, the data discussed in this thesis is generally large and thus **MLE** is the method of choice.

Let $Y_1, ..., Y_L$ be random variables and $y = y_1, ..., y_L$ realizations thereof. Assuming that Y_l are independent and identically distributed random variables with probability function $p(Y_l = y_l \mid \theta)$, the likelihood function of θ is

$$\mathcal{L}_y(\theta) = \prod_{l=1}^{L} P(Y_l = y_l \mid \theta) = P(Y = y \mid \theta). \tag{3.4}$$

The maximum likelihood estimate is then

$$\widehat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}_y(\theta). \tag{3.5}$$

Typically, the maximum likelihood estimate can be determined by computation of the roots of the gradient of $\mathcal{L}_y(\theta)$. In practice, however, the analytical derivation of $\mathcal{L}_y(\theta)$ might be complex. Then, we rely on methods that iteratively approach $\widehat{\theta}_{ML}$, such as the **expectation-maximization (EM)** algorithm.

3.1.3 EXPECTATION MAXIMIZATION

Let us assume that not all variables of a model are observed and that there is a set of hidden data x being drawn from a discrete random variable X with the probability distribution $p(X = x \mid y, \theta)$. We can then define the likelihood function for the complete data:

$$\mathcal{L}_x(\theta) = P_\theta(X = x, Y = y). \tag{3.6}$$

We now want to find the maximum likelihood estimate using **Equation 3.5** on the likelihood function for the complete data. We can replace the likelihood function with the log-likelihood because the logarithm is a monotonic function and thus order preserving. Hence, the parameters optimizing $\sum_x \log \mathcal{L}_x(\theta)$ also optimize $\sum_x \mathcal{L}_x(\theta)$. This is beneficial for analytical and numerical reasons, i.e. preventing floating point underflow.

The **EM** algorithm proceeds in two steps. The **Expectation step (E step)** determines the expected value of the log likelihood of the complete data given the observed data and the current model parameters $\theta^{(t-1)}$. The expected value of a discrete random variable is the weighted average of its possible values, hence:

$$Q\left(\theta, \theta^{(t-1)}\right) = \sum_{x \in Z^L} \log \mathcal{L}_x(\theta)\, P_{\theta^{(t-1)}}(X = x \mid Y = y) \tag{3.7}$$

The **Maximization step (M step)** updates the parameters such that the expected value from the **E step** is maximized:

$$\theta^{(t)} = \arg\max_{\theta} Q\left(\theta, \theta^{(t-1)}\right) \tag{3.8}$$

These two steps are repeated iteratively until convergence, i.e. when $\mathcal{L}_x\left(\theta^{(t)}\right) - \mathcal{L}_x\left(\theta^{(t-1)}\right) < \tau$ with τ being a predefined threshold. Note that the **EM** algorithm converges monotonically to a local maximum. It is important to incorporate previous knowledge into the initial parameter estimation in order to achieve optimal results.

In the next Subsection I will introduce the **Hidden Markov model (HMM)**, a probabilistic framework where a special case of the **EM** algorithm is employed to learn the model parameters.

3.1.4 HIDDEN MARKOV MODELS

HMMs are used to model a sequence of observations emitted by a sequence of underlying *hidden* states. After **HMMs** were first described in the 1960s and initially used for speech recognition problems in the 1980s [151], they are now widely applied to numerous problems in various fields. They are especially popular in bioinformatics and have been used for many tasks including locating genes, detecting CpG islands and segmenting genomes into functional units [152–155]. In this Subsection I will describe the parameterization of **HMMs**, how to compute the probability of the data when the model parameters are known, how to learn the parameters if they are not known, and how to decode the sequence of hidden states for a given model.

PARAMETERIZATION Standard **HMMs** have only one hyperparameter: the total number of discrete hidden states N. Let the set of states be denoted as $Z = \{1, \ldots, N\}$, and let X_l be a discrete random variable with N possible values denoting the hidden state at position $l \in \{1, ..., L\}$. A first-order **HMM** assumes that at any position l in the sequence, the probability of a hidden state only depends on the state at position $l-1$ and is independent of any other state in the sequence or the actual position l:

$$P\left(X_l = x_l \mid X_1 = x_1, \ldots, X_{l-1} = x_{l-1}\right) = P\left(X_l = x_l \mid X_{l-1} = x_{l-1}\right) \tag{3.9}$$

This characteristic is called the Markov property. Hence, we can define the transition probability matrix as follows:

$$A = \{a_{ij}\}, \text{ with}$$
$$a_{ij} = P(X_{l+1} = j \mid X_l = i)$$

For the first position $l = 1$, the probability of state i is given by

$$\pi_i = P(X_1 = i),$$

with π denoting a vector of length N comprising the initial probabilities for every state.

Further, the **HMM** relies on the assumption that an observation only depends on the underlying hidden state and is independent of the actual position l, allowing us to define the emission probability matrix as follows:

$$B = \{b_i(y)\}, \text{ with}$$
$$b_i(y_l) = P(Y_l = y_l \mid X_l = i)$$

Given the model parameters $\theta = \{\pi, A, B\}$, the joint probability of the observed data and a particular sequence of states is

$$P_\theta(Y = y, X = x) = \pi_{x_1} b_{x_1}(y_1) \prod_{l=1}^{L-1} a_{x_l x_{l+1}} b_{x_{l+1}}(y_{l+1}). \tag{3.10}$$

PROBABILITY OF THE DATA UNDER A GIVEN MODEL If we want to infer the marginal probability of the observed data under a given model, we have to integrate the joint probability (**Equation 3.10**) over all possible state sequences. Since the number of possible state sequences is N^L, this can easily become computationally expensive. Especially in the case of genomic data where L scales with the length of the genome, this approach becomes infeasible. However, the use of a dynamic programming approach called the *forward-backward* algorithm allows us to compute the probability of the observed data much more efficiently [154, 155]. The forward process recursively determines the probability of being in state j at position l after having made the current and all previous observations $y_{1:l}$:

$$\begin{aligned} \alpha_j(l) &= P_\theta(Y_{1:l} = y_{1:l} \mid X_l = j) = b_j(y_l) \sum_{i=1}^{N} \alpha_i(l-1) a_{ij} \\ \alpha_j(1) &= \pi_j b_j(y_1) \end{aligned} \tag{3.11}$$

The backward procedure works similarly but from the other end of the sequence. It recursively determines the probability of being in state i at position l and all following observations $y_{l+1:L}$:

$$\begin{aligned} \beta_i(l) &= P_\theta(Y_{l+1:L} = y_{l+1:L} \mid X_l = i) = \sum_{j=1}^{N} a_{ij} b_j(y_{l+1})\ \beta_j(l+1) \\ \beta_i(L) &= 1 \end{aligned} \tag{3.12}$$

The probability of the observed data under the model parameters θ is then given by

$$\begin{aligned} P_\theta(Y = y) &= \sum_{i=1}^{N} \alpha_i(L) \\ &= \sum_{i=1}^{N} \pi_i b_i(y_1)\beta_i(1) \end{aligned} \tag{3.13}$$

LEARNING THE PARAMETERS If the parameters of an **HMM** are not known, they can be learned from the data in an iterative process called the *Baum-Welch* algorithm, a special case of the EM-algorithm [154, 155]. The *Baum-Welch* algorithm aims to yield the maximum likelihood estimate of the parameters given the observed data:

$$\widehat{\theta} = \arg\max_{\theta}\ P_\theta(Y = y)$$

Initially, the parameters have to be guessed and are often assumed uniform, but a more educated guess due to prior knowledge might help the algorithm to converge faster and increases the chance to find the global rather than a local maximum.

In the expectation step, the expected log likelihood is obtained by plugging **Equation 3.10** into **Equation 3.7**. By setting $\rho_{\theta^{(t-1)}}(x) = P_{\theta^{(t-1)}}(X = x \mid Y = y)$ we get:

$$\begin{aligned} Q\left(\theta, \theta^{(t-1)}\right) &= \sum_{x \in Z^L} \rho_{\theta^{(t-1)}}(x) \log P_\theta(X = x, Y = y) \\ &= \sum_{x \in Z^L} \rho_{\theta^{(t-1)}}(x) \log \left(\pi_{x_1} b_{x_1}(y_1) \prod_{l=1}^{L-1} a_{x_l x_{l+1}} b_{x_{l+1}}(y_{l+1}) \right) \\ &= \sum_{x \in Z^L} \rho_{\theta^{(t-1)}}(x) \left(\log \pi_{x_1} + \sum_{l=1}^{L-1} \log a_{x_l x_{l+1}} + \sum_{l=1}^{L} \log b_{x_l}(y_l) \right) \end{aligned} \tag{3.14}$$

In the maximization step, the three summands of $Q\left(\theta, \theta^{(t-1)}\right)$ can be optimized separately. The first summand can be reduced as we sum over all possible state sequences,

but only ever consider the first state:

$$\begin{aligned}\sum_{x \in Z^L} p_{\theta^{(t-1)}}(x)\ \log \pi_{x_1} &= \sum_{i=1}^{N} P_{\theta^{(t-1)}}(X_1 = i \mid Y = y) \log \pi_i \\ &\quad \cdot \underbrace{\sum_{x_2 \in X} \cdots \sum_{x_n \in X} P\left(X_2 = x_2, \ldots, X_n = x_n \mid X_1 = x_1, Y = y, \theta\right)}_{=1} \\ &= \sum_{i=1}^{N} \gamma_i(1) \log \pi_i\end{aligned} \tag{3.15}$$

where $\gamma_i(l)$ denotes the posterior probability of being in state i at position l given the observed data and the current parameters. For that, we use the previously introduced *forward-backward* algorithm:

$$\begin{aligned}\gamma_i(l) &= P_{\theta^{(t-1)}}(X_l = i \mid Y = y) \\ &= \frac{\alpha_i(l)\beta_i(l)}{\sum_{j=1}^{N} \alpha_j(l)\beta_j(l)}\end{aligned} \tag{3.16}$$

Computing the roots of the partial derivative with respect to π_i and subject to $\sum_i \pi_i = 1$, we get:

$$\widehat{\pi}_i = \gamma_i(1). \tag{3.17}$$

The second summand can be reduced accordingly:

$$\begin{aligned}\sum_{x \in Z^L} p_{\theta^{(t-1)}}(x) \sum_{l=1}^{L-1} \log a_{x_l x_{l+1}} &= \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{l=1}^{L-1} P_{\theta^{(t-1)}}\left(X_l = i, X_{l+1} = j \mid Y = y\right) \log a_{ij} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{l=1}^{L-1} \xi_{ij}(l) \log a_{ij}\end{aligned} \tag{3.18}$$

where $\xi_{ij}(l)$ denotes the posterior probability of being in state i and j at position l and $l+1$, respectively, given the observed data and the current parameters:

$$\begin{aligned}\xi_{ij}(l) &= P_{\theta^{(t-1)}}(X_l = i, X_{l+1} = j \mid Y = y) \\ &= \frac{\alpha_i(l)a_{ij}b_j(y_{l+1})\beta_j(l+1)}{\sum_{r=1}^{N}\sum_{s=1}^{N}\alpha_r(l)a_{rs}b_s(y_{l+1})\beta_s(l+1)}.\end{aligned} \tag{3.19}$$

We can again determine the maxima of the partial derivative with respect to a_{ij} and subject to $\sum_j a_{ij} = 1$ and get

$$\widehat{a}_{ij} = \frac{\sum_{l=1}^{L-1}\xi_{ij}(l)}{\sum_{l=1}^{L-1}\gamma_i(l)} \tag{3.20}$$

Finally, the third term becomes

$$\begin{aligned}\sum_{x\in Z^L}\rho_{\theta^{(t-1)}}(x)\sum_{l=1}^{L}\log b_{x_l}(y_l) &= \sum_{i=1}^{N}\sum_{l=1}^{L}P\left(X_l = i \mid Y = y, \theta\right)\log b_i(y_l) \\ &= \sum_{i=1}^{N}\sum_{l=1}^{L}\gamma_i(l)\log b_i(y_l),\end{aligned} \tag{3.21}$$

and maximizing the function with respect to b_i and subject to $\sum_i b_i(k) = 1$ results in

$$\widehat{b}_i(k) = \frac{\sum_{l=1}^{L}\gamma_i(l)\,\mathbb{1}(k = y_l)}{\sum_{l=1}^{L}\gamma_i(l)}, \tag{3.22}$$

where $\mathbb{1}(k = y_l)$ denotes the indicator function

$$\mathbb{1}(k = y_l) = \begin{cases}1 & \text{if } k = y_l \\ 0 & \text{otherwise.}\end{cases} \tag{3.23}$$

See **Appendix A.1.1** for a detailed derivation.

With that, a new set of parameters $\theta^{(t)} = \{\widehat{\pi}, \widehat{A}, \widehat{B}\}$ has been calculated based on the previous set $\theta^{(t-1)}$. Repeating this procedure iteratively until convergence results in locally optimized model parameters. Those can furthermore be used to infer the sequence of hidden states that most likely gave rise to the observed data, a process

called decoding. I will discuss prevalent decoding algorithms in a later paragraph.

In the next paragraph I will describe how to use the *Baum-Welch* algorithm to optimize the emission probabilities when they are modeled by the negative binomial and the log-normal distribution.

THE BAUM-WELCH ALGORITHM FOR NEGATIVE BINOMIAL AND LOG-NORMAL EMISSIONS In the previous paragraph I formally described the learning of the parameters of an **HMM**. When using the negative binomial distribution to model the emission probabilities, the optimization problem of **Equation 3.21** becomes

$$\widehat{\mu}_i, \widehat{r}_i = \underset{\mu_i, r_i}{\arg\max} \sum_{l=1}^{L} \gamma_i(l) \log f_{NB}(y_l; \mu_i, r_i), \tag{3.24}$$

Unfortunately, there is no analytical solution to this maximization problem and an optimal solution can only be found numerically [156].
For the log-normal distribution the optimization problem presents itself as follows:

$$\widehat{\mu}_i, \widehat{\sigma}_i = \underset{\mu_i, \sigma_i}{\arg\max} \sum_{l=1}^{L} \gamma_i(l) \log f_{LN}(y_l; \mu_i, \sigma_i), \tag{3.25}$$

Here it is possible to find an analytical solution for the optimization problem (see **Appendix A.1.5** for a detailed derivation):

$$\widehat{\mu}_i = \frac{\sum_{l=1}^{L} \gamma_i(l) \ln y_l}{\sum_{l=1}^{L} \gamma_i(l)} \tag{3.26}$$

and

$$\widehat{\sigma}_i = \sqrt{\frac{\sum_{l=1}^{L} \gamma_i(l) (\ln y_l - \mu_i)^2}{\sum_{l=1}^{L} \gamma_i(l)}} \tag{3.27}$$

With that, finding the optimal parameters for fitting the read count distribution with a log-normal distribution becomes an easy task that can be achieved in short computational time.

DECODING THE STATE SEQUENCE In this paragraph I will outline two popular approaches for inferring the underlying sequence of hidden states that most likely produced the observed data under a given model [154, 155].

The posterior decoding produces a state sequence x_p containing the most likely state for each position l:

$$\begin{aligned} x_l^{(P)} &= \arg\max_{i \in Z} \; P_\theta(X_l = i \mid Y = y) \\ &= \arg\max_{i \in Z} \; \gamma_i(l) \end{aligned} \tag{3.28}$$

The *Viterbi decoding* in turn produces the most likely state sequence, i.e. it does not maximize $y_x(l)$ for every l independently, but instead maximizes the probability of the whole sequence:

$$x^{(V)} = \arg\max_{x \in Z^L} \; P(X = x \mid Y = y, \theta) \tag{3.29}$$

Instead of computing the probability of every possible path, the *Viterbi* algorithm uses dynamic programming to obtain the same result recursively [154, 155]:

$$\begin{aligned} V_j(l) &= \max_{i \in Z} \left[a_{ij} V_i(l-1) \right] b_j(y_l), \\ V_j(1) &= \pi_j b_j(y_l). \end{aligned} \tag{3.30}$$

If the model contains transition probabilities that are zero, posterior decoding could potentially yield forbidden transitions. In cases where this is undesirable it is crucial to use *Viterbi decoding* in order to preserve the model's grammar. I will give an applied example in **Subsection 4.2.7**, and show how the *Viterbi* algorithm can be used for parameter training in **Subsection 4.2.4**.

3.1.5 EXTENDED HIDDEN MARKOV MODELS

In addition to the standard **HMM**, a plethora of modified and extended topologies of **HMMs** have been described. For example, the factorial **HMM** is based on multiple independent Markov chains, and the tree structured **HMM** is a further extension thereof, which introduces coupling of state variables in a single time step [157]. In triplet Markov models, the distribution of the hidden and observable process is the marginal distribution of a Markov process (X, U, Y), where U is an auxiliary underlying process [158]. Switching state-space models combine real-valued and discrete hidden states in order to model time series with continuous dynamics [159]. Other extensions allow for an infinite or unknown number of states, e.g. the infinite **HMM**

[160] and the hierarchical Dirichlet process **HMM** [161]. Standard **HMMs** do not require class-labeled data but rather learn patterns from unlabeled data, i.e. they fall into the category of unsupervised machine learning. In contrast, supervised **HMMs** have been described that make use of labeled data for parameter learning [162]. I will introduce an application of supervised **HMMs** including additional constraints on the Markov chain topology in **Subsection 4.2.2**.

3.2 GRAPH THEORY

Graph theory describes mathematical structures representing sets of pairwise relations between objects. A graph is a diagram of vertices that are connected by edges, describing their relationship. One of the earliest formal descriptions of a graph goes back to 1736 and the *Seven Bridges of Königsberg*, when the Swiss mathematician Leonhard Euler studied the system of bridges connecting mainland and islands across the river Pregolya in Königsberg (today Kaliningrad, Russia) [163]. He asked whether it was possible to walk over every bridge exactly once with starting and ending at the same place. For that, he reduced a conventional map to the relevant information for this problem which is the land masses and the bridges connecting them, as shown in **Figure 3.1**.

Graphs provide a simple model for many practical problems and are therefore used ubiquitously across all fields of science. In biology, graphs are used extensively to model systems such as gene regulatory networks, biochemical pathways, and many more. I use graphs for the mapping of unalignable genomic regions between distantly related species as discussed in detail in **Section 5.2**.

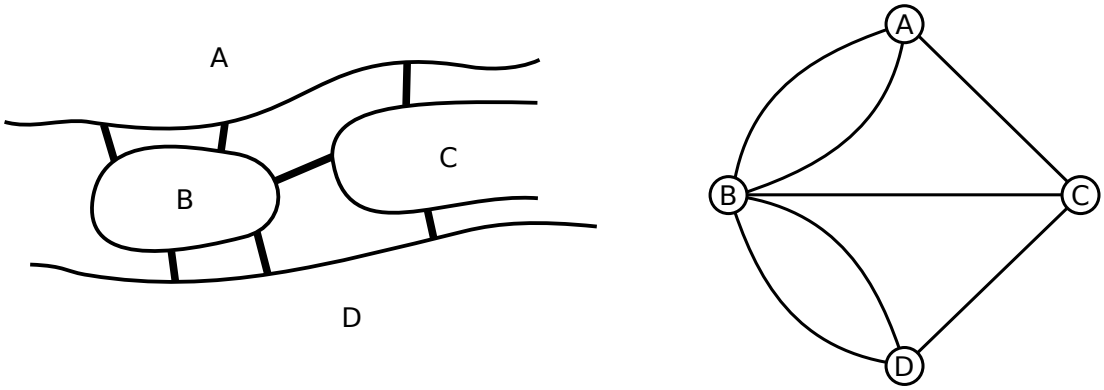

Figure 3.1: A map of the Seven Bridges of Königsberg (left) and its graph representation (right). Adapted from Bóna [163].

3.2.1 FORMAL DEFINITION

A graph G is a pair of sets V and E such that the elements of E are 2-element subsets of V. If G has no loops, i.e. an edge from a vertex to itself, no multiple edges between pairs of vertices, and the edges are bidirectional, then G is a simple undirected graph and can be defined as follows:

$$G = (V, E) \quad \text{such that} \quad E \subseteq \{\, \{x, y\} \mid (x, y) \in V^2 \wedge x \neq y \,\} \tag{3.31}$$

A sequence of distinct edges $e_1 e_2 \dots e_k$ is called a *walk*, and such a sequence with $e_1 = e_k$ is termed a *closed walk*. *Eulerian walks* contain every edge of a graph, and walks with every vertex being present maximally once are called *paths*. The solution to the problem of the Seven Bridges of Königsberg is thus a *closed Eulerian walk*. Graphs with existing paths between every pair of vertices are called *connected*, and if a graph contains all possible edges then it is a *complete* graph. The *degree* of a vertex reveals the number of edges connected to it. In *weighted* graphs, edges are not just mere connections but associated with a numerical weight that holds information about the length of the connection. An example is a graph where the vertices are cities and the edges are the distances between them.

3.2.2 SHORTEST PATH PROBLEM

In many real-world problems we are facing the challenge of finding the shortest path from one vertex to another in a weighted graph. For example, navigation systems consist of graphs with places and streets as their vertices and edges, respectively, and ideally such systems suggest the shortest path between a given start and end point. With increasing graph size, finding the shortest path between two vertices empirically by evaluating every possible path quickly becomes an infeasible task. In 1959, the Dutch computer scientist Edsger W. Dijkstra [164] presented an algorithm that solves the problem for graphs with non-negative edge weights. It is generally perceived as a greedy algorithm, however, it has been proposed to be revalued as a dynamic programming successive approximation procedure [165]. The algorithm has been named after its inventor: **Dijkstra's Shortest Path Algorithm (DA)**.

DIJKSTRA'S SHORTEST PATH ALGORITHM Let $G = (V, E)$ denote a simple undirected graph with non-negative edge weights, i.e. $d(e_i) \geqslant 0 \;\forall\; e_i \in E$. We want to find the shortest path from vertex s to vertex t with $s \neq t$. Let $\delta(v)$ denote the length

of the tentatively shortest path from s to v. Let V split into S and T, where S is the set with all vertices for which we currently have a tentatively shortest path from s, and T is the set of vertices for which we do not. The algorithm starts with the following setting:

$$\begin{aligned} S &= \{s\} \\ T &= V - s \\ \delta(s) &= 0 \\ \delta(v) &= \infty \ \forall \ v \in T. \end{aligned} \tag{3.32}$$

Set the current vertex $c = s$ and evaluate every edge cv with $v \in T$. If $\delta(c) + d(c, v) < \delta(v)$, this means that the path through c to v is shorter than any previously calculated path to v. Set $\delta(v) = \min[\delta(v), \delta(c) + d(c, v)]$. From all evaluated v, choose the vertex with the shortest tentative path, i.e. $v' = \arg\min_v \delta(v)$, move v' to S and set $c = v'$. Continue iteratively until $v' = t$. All tentative shortest paths to any v are stored recursively by defining a pointer $p(v) = c$ as $\delta(v)$ is updated. The final shortest path to t is then found recursively starting at $p(t)$. The time complexity of **DA** has an upper bound of $O(|V|^2)$, but can be decreased to $O(|E| + |V| \log |V|)$ when using priority queues [166].

4 PREDICTION OF CIS-REGULATORY ELEMENTS

4.1 MOTIVATION

Enhancers are **DNA** sequences that act as cis-regulators on the transcription of genes and are therefore at the center of research of transcriptional regulation. The research community is and has been interested in many aspects of enhancer properties and has been asking manifold questions about their role during development [167], how they emerge, become fixed, conserved or eventually disappear over time [168], their mechanistic function and how they target gene promoters [169, 170], their role in disease [167] and many more. Identifying enhancers forms the foundation for all of those questions. There are multiple reasons why this is a difficult task. First, enhancers are relatively small - an ordinary enhancer is thought to comprise a handful of **TFBSs** on a stretch of **DNA** no longer than a few hundred base pairs [171]. Second, the search space is large as enhancers distribute over the whole genome - in case of mammals this is in the order of billions of base pairs with an average of 3.13 **giga base pairs (Gbp)** [172]. Moreover, enhancers do not necessarily target the closest promoter but can regulate genes that are far away, in the renowned case of the *Shh* gene and its ZRS enhancer their linear distance amounts to ~1 **Mbp** [18]. Enhancers can reside up- or downstream of a target gene, in intergenic regions as well as in introns of other genes or the target gene itself. Even exons have been reported to harbor enhancers [173, 174]. They may target one or multiple genes and their activity as well as their target might change based on the tissue, the developmental stage and external stimuli [21].

In recent years, several experimental assays have been designed for the detection of enhancers and for measuring their activity. Classical reporter assays link candidate enhancer sequences to a minimal promoter and a reporter gene in a vector and aim to measure the ability of a **DNA** sequence to enhance the transcription of a reporter gene [175]. Visel et al. [176] provide a resource of non-coding genomic elements with experimentally validated enhancer activity using transgenic mice in their VISTA enhancer browser. These experimental assays test one sequence at a time and are thus not suitable for genome-wide screenings. **Massively parallel reporter assays (MPRAs)** are designed to upscale classical assays using high-throughput technologies and include CRE-seq [177], STARR-seq [178], TRIP [179], FIREWACh [180] and SIF-seq [181]. **MPRAs** have several shortcomings: while some are episomal, meaning that they are oblivious to genomic and chromosomal context, others suffer from low resolution or are not quantitative. For a review, see [182].

Experimental enhancer detection requires the appropriate infrastructure, funding, experienced personnel and time. Thus, there is a demand for replacing these experiments with computational techniques to predict enhancers. An apparent approach is to predict enhancers based on their **DNA** sequence. However, this turns out to be difficult. Existing methods often rely on computationally expensive Deep Learning algorithms and their predictive power is typically limited, especially for predicting the tissues or cell types in which the enhancer is active [183–186]. Conversely, we are even further from being able to design enhancer sequences with desired activity in particular tissues or cell types [187].

It is therefore crucial to explore other properties of enhancers beside **DNA** sequence, i.e. epigenetic features instead of genetic features. The minimal functional unit of an enhancer needs to be able to target a promoter. This happens through the concerted binding of **TFs**, which in turn requires the enhancer **DNA** to be accessible. Chromatin accessibility can be experimentally measured with various assays, for example **ATAC-seq**, a simple and time-efficient method [44]. I elaborate experimental techniques for measuring chromatin accessibility in more detail in **Subsection 2.4.2**.

Accessible chromatin means that it is depleted of nucleosomes. Nucleosomes consist of different types of histones, which are subject to post-translational modifications. Especially some well-studied residues in their N-terminal tails are associated with functional subunits of the genome. For example, elevated levels of **trimethylation of histone 3, lysine 36 (H3K36me3)** are associated with actively transcribed gene bodies [48]. Nucleosomes flanking the central accessible regions of active enhancers have also been observed to exhibit characteristic patterns, namely **H3K27ac** and **monomethylation of histone 3, lysine 4 (H3K4me1)** [50, 53, 58]. Poised enhancers are expected to have their **H3K27** trimethylated instead of acetylated (**H3K27me3**) [58]. The epigenetic signatures of active gene promoters resemble those of enhancers as they also consist of a central stretch of accessible **DNA** and flanking **H3K27ac**. However, **lysine 4 at histone 3 (H3K4)** of nucleosomes flanking promoters is expected to be trimethylated rather than monomethylated. Hence, the methylation level of **H3K4** accounts for one of the most distinctive features between enhancers and promoters. Histone modifications are measured with the widely-used technique **ChIP-seq**, in which specific antibodies against particularly modified histones co-immunoprecipitate the **DNA** wrapped around them [137, 138, 143]. I describe this technology in more detail in **Subsection 2.4.3**.

Not only histones, but the **DNA** itself can be post-translationally modified through methylation of the **cytosine** of CpG-dinucleotides (see **Subsection 2.2.3**). For example, **DNA** methylation at promoters is associated with repressed gene expression [188].

Cell type-specific **DNA** hypomethylation patterns were observed to correlate with those of **H3K27ac** and have been proposed to reflect cell type-specific enhancers [65]. **DNA** methylation can be studied with techniques such as **whole-genome bisulfite sequencing (WGBS)**, the downside of which is that it is expensive and labor-intensive [189, 190]. Alternative methods such as **reduced representation bisulfite sequencing (RRBS)** have been developed in order to reduce those drawbacks, particularly by enriching for CpG-dense genomic regions before sequencing and thereby reducing the sequencing cost to approximately 1% compared to **WGBS** [191].

In 2010, Kim et al. [70] described the production of short non-coding transcripts at enhancers. These so-called **eRNAs** potentially emerge as a consequence of the enhancers' spatial proximity to **RNAP II** during transcription initiation at the promoter. It is unclear, however, whether **eRNAs** also have a functional role themselves [70, 192]. Among other experimental techniques that measure the production of **eRNAs**, **CAGE** was used by the **FANTOM** consortium to establish a comprehensive atlas of transcribed enhancers in multiple organisms and cell types [79, 84]. However, not all active enhancers are observed to produce **eRNA**. In mouse cortical neurons, **eRNAs** have been detected at approximately 25% of all enhancers [70].

Evolutionary conservation of **DNA** sequence might indicate its functional importance [85, 86]. Naturally, this has led to the selection of conserved elements as candidate enhancer regions [176, 193–195]. While highly conserved enhancers are associated with the regulation of fundamental processes such as embryonic development, recently evolved and species-specific enhancers that contribute to positively selected alterations in gene expression are consistently overlooked by that strategy [93, 97]. Moreover, enhancers have been shown to exhibit functional conservation beyond the sequence level [98, 99, 101].

Taken together, enhancers can be described by genomic and epigenomic features, are frequently transcribed and often evolutionarily conserved. However, no single feature is sufficient to identify enhancers genome-wide and as a consequence, the exact definition of an enhancer is still debated. There exists a plethora of methods that implement respective subsets of the described features to predict enhancers. They can be divided in two broad classes - supervised and unsupervised methods. Unsupervised methods do not require labeled training sets. They directly learn patterns on the provided data. This entails that they do not include any prior biological knowledge, e.g. about the molecular structure of enhancers. Moreover, the results of unsupervised methods need to be inspected and interpreted by the user. For example, unsupervised **HMMs** yield transition and emission parameters which can be evaluated and assigned to a type of genomic element. This process inherently varies on an indi-

vidual basis and can lead to bias. Supervised methods rely on a labeled training set in order to learn how to distinguish between positive and negative examples. Many mathematical models have been employed in both an unsupervised and a supervised manner (see [196, 197] for review). One of the most prominent ones is the **HMM** [151]. **HMMs** can be used to infer an unknown state associated with each position in a given sequence of observations. They assume that observations are generated by an underlying hidden state emitting symbols according to a particular probability distribution. **HMMs** are therefore ideal for the task of recognizing chromatin states based on the observed sequence of histone modification patterns, and have been used repeatedly for that purpose in an unsupervised, as well as a supervised fashion. Chromatin annotation methods such as ChromHMM [198], **EpiCSeg** [199] or Genostan [200] implement an unsupervised **HMM**, i.e. the main hyperparameter is the desired number of states. These methods require the user to interpret and annotate the learned states based on previous knowledge about functional elements in the genome, e.g. that promoters are enriched in H3K4me3 signal. Won et al. [201] turn this approach around and use supervised **HMMs** with a left-right structure to predict different genomic modules such as enhancers and promoters, and incorporate the modules into one model. They integrate existing knowledge into the model by learning the parameters on pre-selected training sets. However, their model allows the modules to be passed through in many different ways, e.g. skipping the state representing the nucleosome-free region where transcription factors can bind, leaving the method very sensitive for detecting false positives. Other methods rely on different mathematical models in order to predict enhancers [202–204], and many of them do not consider prior biological knowledge about enhancers such as their diverse lengths or their heterogeneous molecular structure. Others include great numbers of features and are therefore very data expensive, making it difficult to study particular cell types for which it is infeasible to acquire a lot of experimental data. [205].

To address this, I designed **enhancer Hidden Markov Model (eHMM)**, which implements multivariate modeling of functional genomics data in a supervised **HMM**. **eHMM** comprises enhancer and promoter modules that capture the physical structure of gene regulatory elements, i.e. a central accessible stretch of **DNA** flanked by nucleosomes that exhibit typical **histone modifications**. In the following Sections I describe the method, assess the performance of **eHMM** within and across developmental stages and cell types, compare it to both unsupervised and supervised methods and show how **eHMM** raises the benchmark of current enhancer prediction in terms of prediction accuracy, spatial resolution and low levels of false-positives. Based on measuring the area under the precision-recall curve, **eHMM** outperforms previous methods. Moreover, **eHMM** is easy to interpret, yields predictions with a high resolution and provides a pre-trained model that can robustly be applied across samples.

The advantages of **eHMM** come at a price. The method is rather conservative, typically identifying small numbers of enhancers with high confidence in the order of a few thousands genome-wide. Moreover, its promoter module mainly serves the purpose of increasing specificity of the enhancer module by competition. It is therefore not optimized for promoter identification and for example oblivious to asymmetrical promoters. In the remainder of this Section I will describe in detail how **eHMM** uses a supervised **HMM** with a constrained underlying Markov chain, incorporates prior biological knowledge about the molecular structure of enhancers in a dynamic model to predict heterogeneous enhancers of variable sizes on the basis of a minimal set of features.

4.2 METHODS

There are four main aspects that I incorporated into the development of **eHMM**. First, I set out to include as few, non-redundant and commonly available or cheaply producible features as possible. These turned out to be a chromatin accessibility assay (e.g. **ATAC-seq** or **DNase-seq**), **H3K27ac**, **H3K4me1** and **H3K4me3**. I will discuss the selection of features in more detail in **Subsection 4.2.1**. For the sake of readability I will consistently use **ATAC-seq** in the subsequent parts of this Chapter when referring to any chromatin accessibility assay.

Second, the method was supposed to capture the molecular structure of enhancers. In other words, it should be able to distinguish accessible chromatin from nucleosome-occupied chromatin, and it should identify an enhancer only if those chromatin states are arranged in a meaningful order, i.e. a central nucleosome-free region flanked by nucleosomes. To that end, I decided to use a **hidden Markov model**, but instead of a standard implementation, I set out to constrain the **HMM** in terms of transition probabilities. I elaborate on this procedure in **Subsection 4.2.2**.

Third, instead of using the **HMM** in its conventional unsupervised fashion, I aimed to supervise model training in order to incorporate prior biological knowledge into the model and therefore improving prediction quality. In addition, supervised models avoid potential ambiguity of user-interpreted results from which unsupervised models may suffer. The model consists of the combination of three supervised models representing enhancers, promoters and background, each being trained individually on a designated training set. As promoters and enhancers exhibit a substantial overlap in histone modification patterns, this distinction helps the enhancer module not to primarily detect annotated promoters. I acknowledge reports attributing enhancer function to some promoters [206], however, this dual role is not within the scope of

this thesis. For clarity, I will refer to the three models as *modules* and reserve the name *model* for the full model, i.e. the combination of the modules. I describe the definition of the training sets and model training of the three individual modules in **Subsections 4.2.3** and **4.2.4**, respectively. **Subsection 4.2.5** dwells on model architecture, i.e. how the individually trained modules are combined into one model.

Lastly, I intended to develop a method that is widely applicable, i.e. to any cell type without the need of training a separate model each time. To that end, **eHMM** implements a pre-trained model including a normalization step for users to apply to any data set. I address model implementation and data normalization in **Subsections 4.2.9** and **4.2.10**, respectively.

4.2.1 FEATURES

eHMM implements enhancer and promoter modules, subsequently referred to as the foreground modules, designed to capture an enhancer's or promoter's topology, respectively. These topologies consist of a central accessible stretch of **DNA** flanked by two nucleosomes. Chromatin accessibility is measured with assays such as **ATAC-seq** or **DNase-seq**. Nucleosomes are detected from the occurrence of **ChIP-seq** signals for the three histone modifications **H3K27ac**, **H3K4me1** and **trimethylation of histone 3, lysine 4 (H3K4me3)**. **H3K27ac** is generally associated with active chromatin, whereas ratios of **H3K4me1** over **H3K4me3** are typically high at enhancers and low at promoters. This small set of four features provides a maximal amount of information while being minimally redundant at the same time. Moreover, it consists of only the most prevalent histone marks for which antibodies are available for many species, tissues and developmental stages.

4.2.2 A SUPERVISED AND CONSTRICTED HIDDEN MARKOV MODEL

Standard **HMMs** are unsupervised, i.e. they learn patterns from unlabeled data. Here I want the model to learn the molecular structure of enhancers and promoters. I thus use a supervised approach to learn these structures from predefined training sets.

The molecular structure of enhancers and promoters is understood to consist of two chromatin states: a central stretch of nucleosome-free, i.e. accessible chromatin (s_A) flanked by nucleosomal chromatin (s_N) to each side. These chromatin states are the hidden states of the **HMM** and their properties are inferred from the observable

functional genomics input data. The foreground modules thus consist of the set of hidden states

$$S = \{s_A, s_N\}$$

and enhancers and promoters are expected to exhibit the minimal state sequence $s_N - s_A - s_N$. In **Subsection 4.2.4** I will discuss that s_A and s_N may themselves be sets of multiple states.

The key characteristic of both foreground modules is directionality, as depicted in the corresponding Markov chain in **Figure 4.1**. Both enhancer (E) and promoter (P) modules can only be reached through transitions from a state of the background module (BG) to states representing the first nucleosome (N1). From there, accessible chromatin states (A) and later a second nucleosome state (N2) have to be visited before returning to the background module. In addition, self-transitions allow the model to capture regulatory elements of variable lengths. I discuss the implementation of the model's directionality constraints in the following **Subsection 4.2.4**.

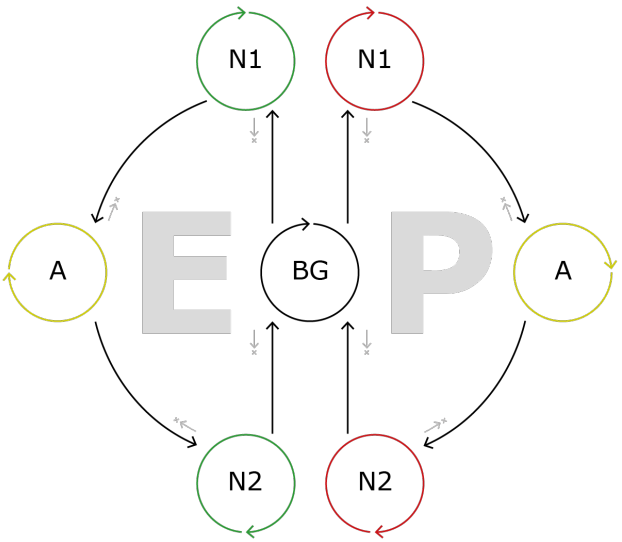

Figure 4.1: Schematic Markov chain of **eHMM**'s underlying constricted Hidden Markov Model. Enhancer and promoter modules (E and P, respectively) consist of states N1, A and N2 which can only be transitioned in a directed fashion from the background module (BG).

4.2.3 TRAINING SETS

In this Subsection I will describe how I assembled the training sets used for training the modules.

ENHANCER TRAINING REGIONS The enhancer module is designed to learn chromatin states from genomic regions that are confidently classified as enhancers. To date, there is no gold standard set of true enhancers. However, there is a wealth of experimental approaches for identifying enhancers [84, 176, 178]. Since the model learns patterns of **ATAC-seq** and histone modification **ChIP-seq** signals, I defined the training set based on criteria independent of these features. As I introduced in **Subsection 2.4.4**, enhancers can be located experimentally using **CAGE** sequencing on RNA samples. FANTOM5 is a project of the **Functional Annotation of the Mammalian Genome (FANTOM)** consortium that provides data sets for multiple tissues in many vertebrate species [84]. I applied the following protocol to the publicly available **CAGE** data sets for mouse **embryonic stem cells (ESC)** of cell line E14, liver E12 and lung E17 in order to define the enhancer training regions: I set a minimal threshold of 11 (**ESC**) and 5 (liver, lung) **CAGE**-tags per region resulting in 5573, 537 and 642 regions, respectively. These regions are expected to include many false positives as a consequence of the low thresholds and necessitate further filtering. I performed k-means clustering on the regions' **ATAC-seq**, **H3K27ac** and **H3K4me1/3 ChIP-seq** signals with $k = 5$ and selected the cluster with the strongest active enhancer signature consisting of 920 regions in **ESC**. The discarded clusters exhibited typical patterns of promoters, poised enhancers, or were depleted of any signal. The model topology requires the training regions to be accurately defined, i.e. to start and end at nucleosome positions. To that end, I used MACS2 [207] with default settings to determine **H3K27ac - ATAC-seq - H3K27ac** peak triplets with a width of less than 2 **kbp** overlapping with the active enhancer regions, followed by the removal of neighboring regions (pairwise distance of less than 2 **kbp**). This procedure resulted in a set of 647 active enhancer regions in **ESC**, from which I randomly sampled 300 regions. In liver and lung, the final number of enhancers in the training set was 118 and 62, respectively.

PROMOTER TRAINING REGIONS I defined the promoter training regions in a similar fashion: **H3K27ac - ATAC-seq - H3K27ac** peak triplets with a width of less than 2 **kbp** belonging to the cluster with the strongest active promoter signature and overlapping an annotated promoter from the **University of California Santa Cruz (UCSC)** knownGene database [208], but not a previously defined enhancer training

region. In order to favor the selection of active promoters, I applied an additional filter for **H3K27ac** levels greater than or equal to the minimum **H3K27ac** level of the previously defined enhancer training regions. From the resulting 3029 regions in **ESC**, I randomly sampled 300 regions to give rise to the training set for the **ESC** promoter module. I obtained the training sets for liver and lung analogous to the described procedure.

BACKGROUND TRAINING REGIONS There are several approaches to train a background module. I will elaborate on them in detail in **Subsection 4.2.4**. One of them is a standard **HMM** learned on a training set representing genomic background, i.e. all types of genomic elements except enhancers and active promoters. To this end, I defined mammalian genomic proportions by roughly approximating the numbers reported for the human genome by Kellis et al. [198]. This resulted in 10% enhancers, 5% active promoters, 5% inactive promoters, 10% genic and 70% intergenic regions. I obtained the training set for the background module by randomly sampling 2 kb genomic regions according to these proportions with respect to **UCSC** knownGene annotations, leaving out regions annotated as enhancers or active promoters.

Figure 4.2 shows the average signal distributions for the enhancer, promoter and background training regions in all three cell types.

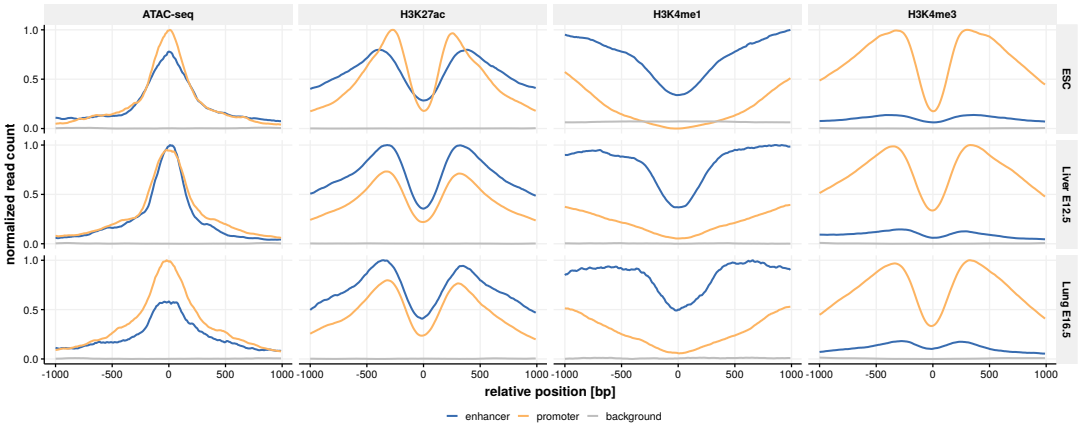

Figure 4.2: Distribution of normalized read counts for training regions of mouse **ESC** E14, mouse embryonic liver E12.5 and mouse embryonic lung E16.5.

4.2.4 MODEL TRAINING

In this Subsection I will discuss the process of individually training the three modules for enhancers, promoters, and background on the previously defined training sets.

TRAINING THE FOREGROUND MODULES The training procedure for the foreground modules comprises a three-step learning process. First, a conventional five-state **HMM** is learned on the respective training sets. I set the number of states to five with the intention of learning multiple hidden states per chromatin condition. This will grant additional flexibility to the model as it will be able to learn different variants of the same functional state, e.g. two states s_{A1} and s_{A2} exhibiting strong and weak chromatin accessibility, respectively.

Second, states are assigned to represent either chromatin accessibility (s_A) or nucleosomes (s_N) based on their emission parameters (see **Subsection 4.2.2** and an example in **Figure 4.4**, left). The automated state selection assigns to $S_A = \{s_A^{(1)}, s_A^{(2)}\}$ the two states with the highest **ATAC-seq** to **H3K27ac** (or **DNase-seq** to **H3K27ac**) ratio. From the remaining three states, the two with the highest (enhancer module) or lowest (promoter module) **H3K4me1** to **H3K4me3** ratio are assigned to $S_N = \{s_N^{(1)}, s_N^{(2)}\}$. The ratios are calculated on the mean of the fitted log-normal distributions of each state. Then, states in S_N are duplicated to $S_{N1} = \{s_{N1}^{(1)}, s_{N1}^{(2)}\}$ and $S_{N2} = \{s_{N2}^{(1)}, s_{N2}^{(2)}\}$ and arranged in a directed order together with the states in S_A. Transitions conflicting with the intended directionality, e.g. from a state in S_{N2} back to a state in S_A, are forbidden by setting the corresponding transition probabilities to zero. See **Figure 4.3** for illustration.

Third, the newly constructed model's parameters are used as initial values for parameter re-estimation. The aim of this step is to modulate the transition probabilities between the rearranged states without changing the emission probabilities and thereby preventing states previously assigned to a particular class to adapt. For that purpose, I use *Viterbi training* [209] instead of the *Baum-Welch* algorithm. *Viterbi training* is a simplification of the *Baum-Welch* algorithm and results in an approximation of the maximum likelihood estimate. Instead of accounting for all possible paths, only the most probable path is considered during parameter re-estimation (see **Equation 3.30**). Moreover, *Viterbi training* allows to force the state sequence to end with a state in S_{N2} for every training sample, maintaining the structural constraints of enhancers and

promoters. For that purpose, the transition matrix for the last step is modified such that only transitions to an end state $j \in S_{N2}$ are possible:

$$\begin{aligned} A^{(L)} &= \{a_{ij}^{(L)}\}, \text{ with} \\ a_{ij}^{(L)} &= P(X_L = j \mid X_{L-1} = i) \\ &= \begin{cases} a_{ij} & \text{if } j \in S_{N2} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4.1}$$

$A^{(L)}$ is then normalized by the sum of its columns such that $\sum_j a_{ij}^{(L)} = 1$. With these constraints I aim to achieve an accurate representation of enhancer and promoter characteristics reflected by both emission and transition parameters.

After the decoding step of each learning iteration, the model parameters π and A are updated according to the principles discussed in **Subsection 3.1.4**, while the emission parameters are not allowed to change:

$$\begin{aligned} \widehat{\pi}_i &= \begin{cases} 1 & \text{if } i = \arg\max_{j \in Z} V_j(1) \\ 0 & \text{otherwise} \end{cases} \\ \widehat{a}_{ij} &= \frac{\sum_{l=1}^{L-1} \delta_i(l)\, \delta_j(l+1)}{\sum_{l=1}^{L-1} \delta_i(l)} \end{aligned} \tag{4.2}$$

where

$$\delta_l(i) = \begin{cases} 1 & \text{if } i = x_l^{(V)} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

TRAINING THE BACKGROUND MODULE There are different possible approaches to train a background module. For example, I can deduce it from the previously trained foreground modules by adopting the number of states and their emission probabilities, but overwriting their transition probabilities such that they are uniform. That way, I emphasize the foreground modules' capability to capture the molecular structure of enhancers and promoters reflected by the transition probabilities. I will call this type of training approach *FGtoBG*.

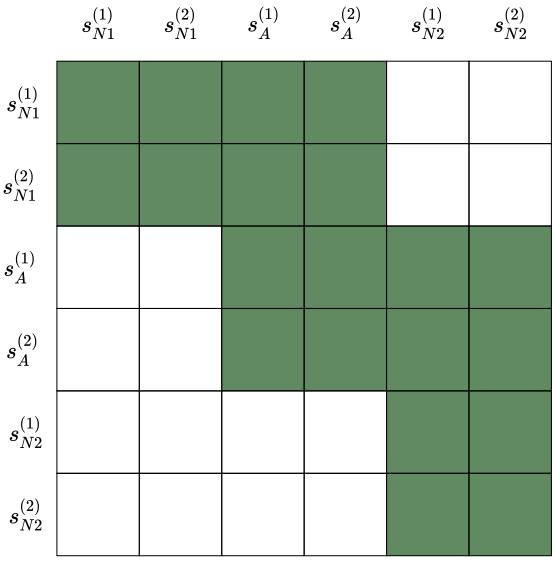

Figure 4.3: Schematic illustration of the transition probabilities of a foreground module. Allowed transitions are green, forbidden transitions are white.

Further, analogous to the foreground training procedure, I can define a training set that represents the aforementioned genomic proportions in mammals, see **Subsection 4.2.3**. Varying the number of states will tell us the impact of that hyperparameter on total model performance. I will call this type of training approach *Standard*.

In addition, I will customize the *Standard* background module by removing states with enhancer- and promoter-like emission patterns. This could have an impact on the sensitivity of the model and thus the overall performance by reducing direct competition for the states in the foreground modules. I would expect such a model to be more sensitive to emission patterns only remotely similar to the foreground states. This type of training approach will be termed *Reduced*.

I will compare the performance of the background modules *Standard*, *Reduced* and *FGtoBG* with varying numbers of states in **Subsection 4.3.1**.

4.2.5 MODULE COMBINATION

Once the three modules are trained, they are combined into one model consisting of all states (see example in **Figure 4.4**). Transitions between states of different modules are either set to zero because they are not allowed, or estimated in the case of $S_{BG} \rightarrow S_{N1}$ or $S_{N2} \rightarrow S_{BG}$ transitions. For the first, I refer to the estimated number of enhancers (399,124) and promoters (70,292) in the human genome as stated by the ENCODE consortium [210], as well as to the total human genome size of roughly

3 billion **bps** according to genome assembly GRCh38 [3], and a bin size of 100 **bps**. These numbers led to estimated $S_{BG} \rightarrow S_{N1}$ transition rates of 1.33% and 0.23% for enhancers and promoters, respectively, and I expect them to be good estimates for other mammalian genomes, too. I set $S_{N2} \rightarrow S_{BG}$ transitions to the learned values of $S_{N1} \rightarrow S_A$ transitions as the size of all nucleosomes is expected to be equal and thus the size distributions modeled by the transition parameters of S_{N1} and S_{N2} should be too.

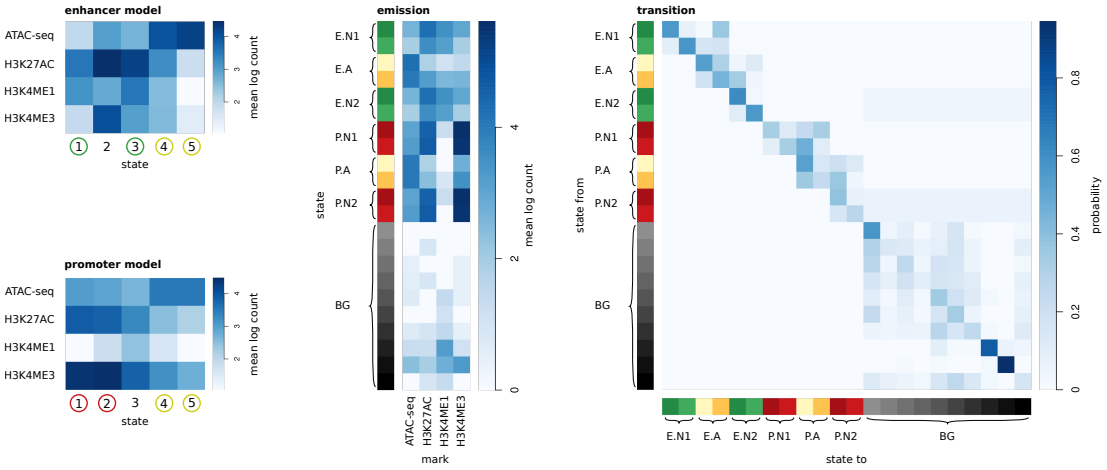

Figure 4.4: Model parameters. Left: state selection based on emission patterns of the foreground modules. Selected states are encircled in green (enhancer nucleosomes), red (promoter nucleosomes), and yellow (accessibility). Right: emission and transition parameters of the full model.

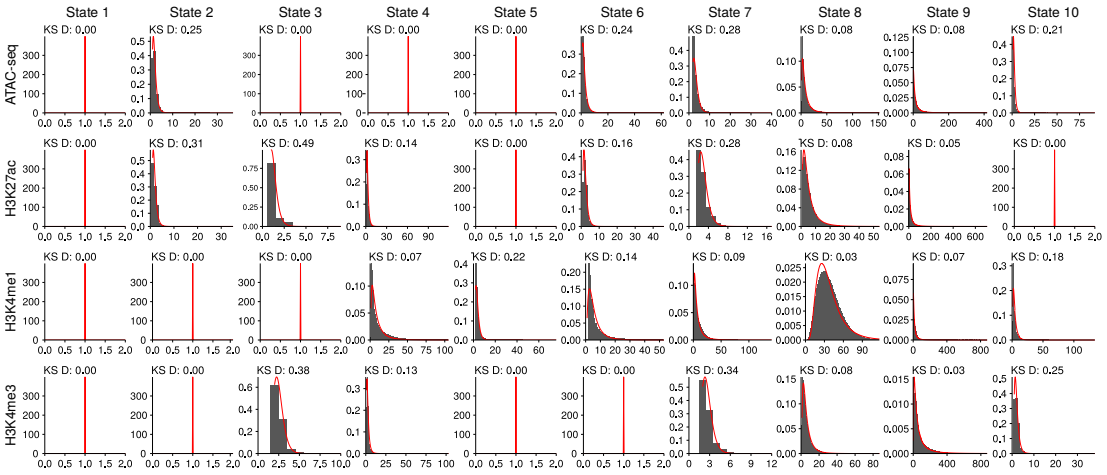

Figure 4.5: Histograms of read count data (grey) and fitted log-normal distributions (red) of a standard 10-state **HMM** learned on whole genome **ESC** data.

4.2.6 EMISSION DISTRIBUTIONS

Mammana et. al [211] show that multivariate read count data can be accurately modeled using the negative multinomial distribution. However, the fitting procedure for negative multinomials requires a complex numerical approximation. Instead, I modeled the read count data with independent log-normal distributions, which appear to be both a better fit for the data as well as the analytical fitting procedure being much easier, see **Subsection 3.1.4**. Fit quality is demonstrated in **Figure 4.5**, showing the read count data and the fitted log-normal distributions in a standard 10-state model learned on whole genome **ESC** data. **Kolmogorov-Smirnov (KS)** distances between the data and the fits were computed for all features and states, ranging from 0.00 to 0.49 with a median of 0.08. Some components model a single coverage value and I assume here that such states have a **KS** distance of 0. In contrast, marginal negative binomial fits show **KS** distances ranging from 0.02 to 0.29 with a median of 0.09 (data not shown).

4.2.7 DECODING AND SCORING

There are several decoding algorithms that yield a state sequence from a learned HMM, see **Subsection 3.1.4**. *Posterior decoding* determines the path with the most probable state at any genomic position. However, it may not preserve the model's grammar, which is essential in order to prevent forbidden transitions e.g. from a state representing an accessible region to a background state. Hence, I use the *Viterbi decoding* algorithm which returns the globally most likely path resulting in a particular number of predicted enhancers without the requirement for finding an optimal prediction threshold. However, while these predictions all belong to the globally most likely path, they might differ in local certainty. The *forward-backward* algorithm provides a posterior probability for the respective state at each position, considering all possible paths. Summing over the posteriors of the enhancer module's states representing accessibility at every position provides a measure of prediction certainty with expected maxima at the center of predicted enhancers. I denote the position-specific enhancer score $s_l^{(E)}$:

$$s_l^{(E)} = \sum_{i \in S_A^{(E)}} P(X_l = i \mid Y = y, \theta) \quad = \sum_{i \in S_A^{(E)}} \gamma_i(l)$$

Figure 4.6 shows an example genomic region classified as an enhancer and the corresponding enhancer score.

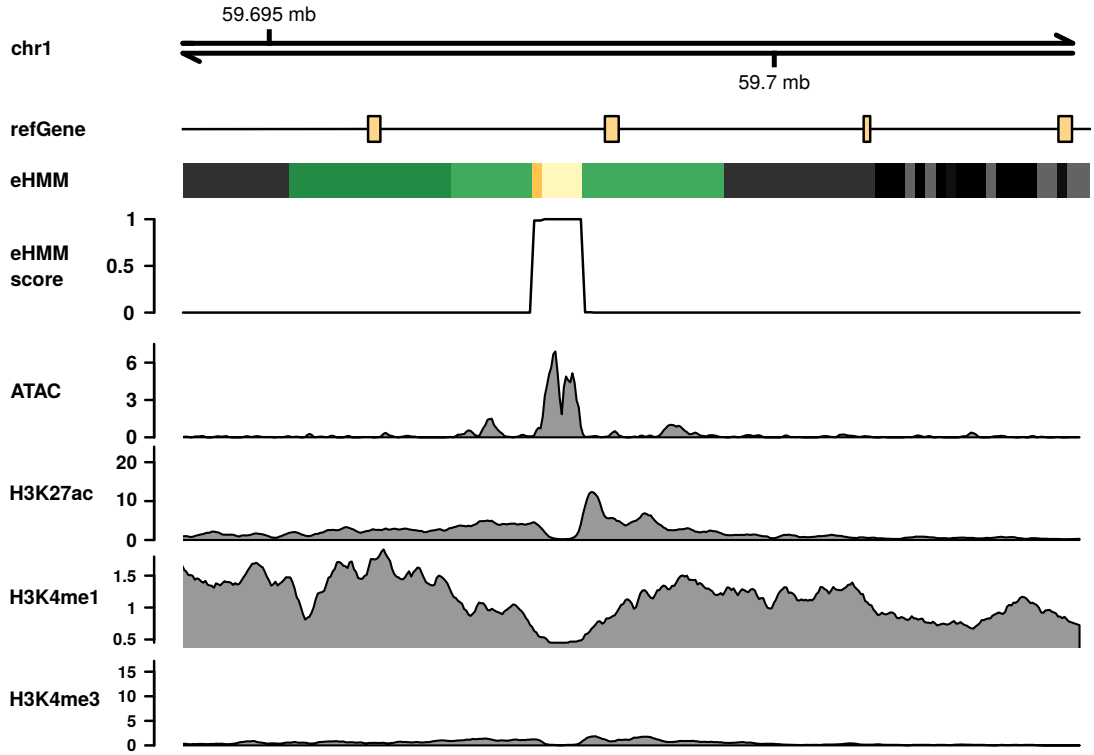

Figure 4.6: Example genomic region with functional genomics data in mouse **ESC** with the segmentation and corresponding enhancer score from **eHMM**. The color code in the segmentation track corresponds to that of **Figure 4.4**.

4.2.8 TESTING

I assessed the method's ability to correctly classify enhancers with different setups. First, I evaluated the method's performance within a certain cell type using cross-validation. Then, I tested the method's predictions in a given cell type when the model was trained in another, referred to as cross sample validation. In the following paragraphs I will outline the definition of the test sets and describe how I evaluate the method's performance within as well as across cell types.

TEST SETS I used the previously described training regions in **ESC**, liver E12.5 and lung E16.5 for within as well as cross cell type validation. In addition, I defined test sets in **ESC**, liver E14.5 and lung E14.5 using regions from the EnhancerAtlas [212]. I processed the data sets by combining regions within 500 **bps**, excluding regions that are located within 2 **kbp** of annotated promoters from the **UCSC** knownGene database and centering on the highest overlapping **ATAC-seq** peak in order to emphasize my intention to focus on functional enhancers. Notably, this led to data set reductions of 68%, 83% and 66% for **ESC**, liver and lung, respectively. I complemented the test sets with randomly sampled regions according to the proportions

of functional elements in mammalian genomes with respect to **UCSC** knownGene annotations, see **Subsection 4.2.3**.

PERFORMANCE EVALUATION In order to assess the method's performance within a certain cell type, I performed a 5-fold cross-validation scheme on the previously described unbalanced training and test sets, such that each test set contained $1/5$ of the original enhancer training set, while the model was trained on the remaining $4/5$. Cross sample validations were obtained by training the model on the respective full training sets and evaluating the labeled test sets. The model's performance was quantified by calculating the **area under the precision recall curve (AUPRC)**. Precision and recall are defined as follows:

TP	True Positives	Enhancers correctly classified as enhancers
FP	False Positives	Non-enhancers falsely classified as enhancers
FN	False Negatives	Enhancers falsely classified as non-enhancers

Precision tells us how many of the regions classified as enhancers are actually true enhancers:

$$\text{precision} = \frac{TP}{TP + FP} \tag{4.4}$$

Recall denotes how many of the actual enhancers were correctly classified as enhancers:

$$\text{recall} = \frac{TP}{TP + FN} \tag{4.5}$$

A very stringent classifier typically reaches high precision, but low recall, whereas the opposite is true for a very lenient classifier. Plotting precision against recall with different prediction thresholds thus gives a good indication on how well the classifier finds a trade-off between those measures. Calculating the **AUPRC** provides a quantification that can be used to compare different methods exposed to the same prediction problem. **Figure 4.7** shows example precision recall curves for different types of classifiers for a balanced data set (i.e. the number of positive examples is half of the total number of examples in the data set). Note that, in contrast to other metrics such as prediction accuracy or **receiver operating characteristic (ROC)**, precision and recall are also suitable for imbalanced data sets. In that case, the baseline of the random classifier would move to the level of the fraction of positive examples in the set.

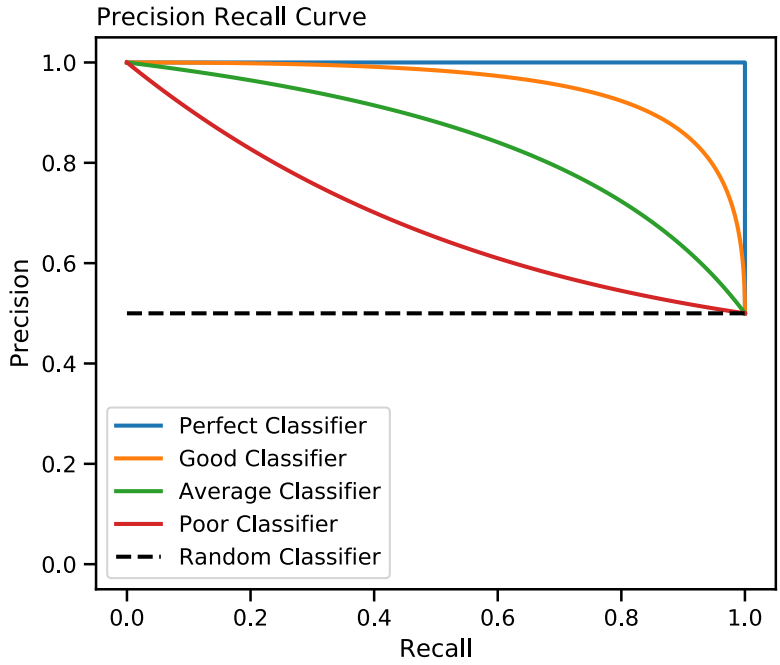

Figure 4.7: Exemplary precision recall curves for five different classifiers on a balanced data set. The area under the curve of a random classifier represents the fraction of positive examples in the data set.

4.2.9 IMPLEMENTATION

I implemented the method as an R software package [213] named **eHMM** available at `https://github.com/tobiaszehnder/ehmm`. In 2015, Mammana et al. developed **EpiCSeg** [211], a tool for genome segmentation using an unsupervised **HMM** on epigenetic features. Since the framework of that method is essentially very similar to the requirements of the method described here, I used it as a basis for building the method. I adopted already implemented basic functionalities such as the initialization algorithm or the *Baum-Welch* algorithm, and extended the framework by developing new functions for tasks and features specific for **eHMM**.

eHMM is implemented primarily in C++, and integrated with R using the Rcpp package [214]. It implements multi-threading for dealing with large sets of data and is available as an interface in R as well as the command line program.

eHMM comprises three subprograms which are supposed to run in sequence: `learnModel`, `constructModel` and `applyModel`. First, `learnModel` learns a standard **HMM** for a given set of genomic regions based on the features described in **Subsection 4.2.1**. The user inputs read count data in bam file format for the required features and specifies the desired number of hidden states. This step is supposed to be exerted separately for previously defined training sets for enhancers, promoters and background re-

gions. Once these three models are learned, `constructModel` is applied to combine the modules to one big model as described in **Subsection 4.2.5**. Lastly, `applyModel` takes the learned model and applies it on a given set of genomic regions, e.g. a labeled test set for model testing or the whole genome.

Defining labeled training sets takes time and access to appropriate data. I thus decided to equip the software package with a pre-trained model that spares the user the potentially laborious first two steps. I trained this model on mouse **ESC** E14 data. The determination of the training sets is described in **Subsection 4.2.3**. If the user chooses to use the provided pre-trained model, the query data will be normalized to the data that was used during model training using quantile normalization. I will discuss quantile normalization in **Subsection 4.2.10**.

4.2.10 QUANTILE NORMALIZATION

Data from different ChIP-Seq experiments may vary in their total number of reads and their read count distributions may be scaled differently. Therefore, in order to apply a model learned on a specific cell type to another cell type, input data has to be brought to the same scale. For that, I used quantile normalization to adjust the statistical properties of a query distribution (the data the model is applied to) to a reference distribution (the data the model was learned on) [215]. This method minimizes the distance between the query and reference cumulative distributions by an order-preserving rescaling of the query count values. In practice, it orders a matrix containing read counts with rows representing genomic locations and columns representing the query and reference samples in a column-wise fashion. Then, values from the reference column are assigned to the query column before the columns are re-ordered back to the original order. We can assess the effect of the normalization procedure by comparing MA plots of the unadjusted and the normalized data. MA plots show the two metrics M and A. M is the binary logarithm of the count ratio between the reference (R) and the query (Q) samples ($M = \log_2 R/Q$), and A is the average log count ($A = \frac{1}{2}\log_2 RQ$). Hence, MA plots visualize intensity-dependent differences between samples. **Figure 4.8** depicts exemplary MA plots and read count distributions from **H3K4me1 ChIP-seq** experiments in mouse **ESC** E14 and mouse embryonic lung E16.5. The data shown is a random sample of 20000 genomic locations on chromosome 1. Quantile normalization eliminates intensity-dependent differences between the two samples and leads to a strong overlap between the two distributions.

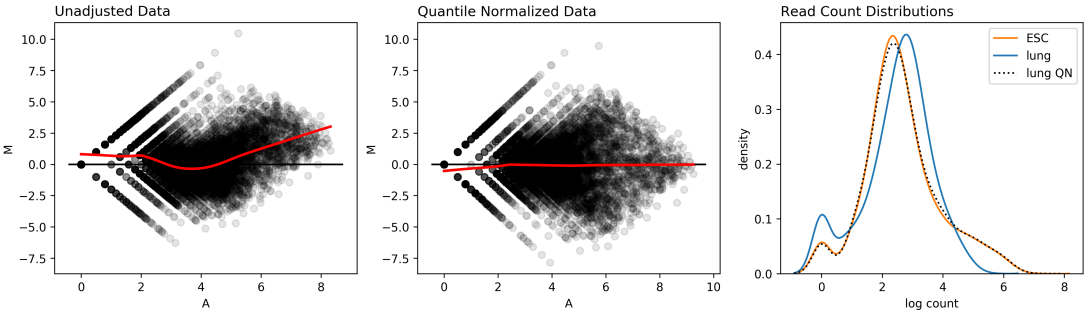

Figure 4.8: Effect of quantile normalization on read count data. Left panel: MA plot of unadjusted data. Middle panel: MA plot of quantile normalized data. The respective red lines show local regression (LOWESS). Right panel: Log count distributions in different cell types before and after normalization.

4.2.11 DATA

DATA SOURCES I investigated five specific mouse samples regarding cell type and developmental stage: **ESC**, embryonic liver E12.5 and E14.5 as well as embryonic lung E14.5 and E16.5. **ATAC-seq** and **HM ChIP-seq** data from liver and lung samples were obtained from **ENCODE** [210]. I downloaded **ESC HM** and **TF ChIP-seq** as well as **Methylated DNA immunoprecipitation followed by sequencing (MeDIP-seq)** data from the **NCBI Gene Expression Omnibus (GEO)** [216], and converted genome coordinates from mm9 to mm10 with CrossMap [217]. I obtained sequence conservation data from phastCons conservation scores from UCSC [218]. An overview of all used data and their accession numbers is given in **Appendix A.2.1**.

DATA PROCESSING I downloaded the raw data fastq files using the **SRA** toolkit [219] and processed fastq to bam files using the **Burrows-Wheeler Alignment (BWA)** tool [220] for mapping and SAMtools [221] for filtering, sorting and removing duplicates. **eHMM** implements the algorithm bamsignals [199] to calculate read counts for bins with a width of 100 **bps**. In order to estimate the fragment centers and with an expected fragment length of 150 **bps**, bamsignals adds a default shift of 75 **bps** to **ChIP-seq** reads. In contrast, chromatin accessibility assays are treated with a shift of zero as the interest of these experiments lies on the actual cutting sites. I added a pseudo-count of 1 to prevent taking logarithms of entries with value zero.

4.3 RESULTS

In this Section I will assess the quantitative and qualitative performance of **eHMM**. I will discuss different strategies for defining the background module (**Subsection 4.3.1**), evaluate the prediction performance within and across cell types and data sources (**Subsections 4.3.2** and **4.3.3**) and compare the performance of **eHMM** to state-of-the-art methods in terms of precision and recall (**Subsection 4.3.4**) as well as run time (**Subsection 4.3.8**). Finally, I will discuss the properties of genome-wide predicted enhancers on a qualitative level (**Subsections 4.3.5** to **4.3.7**).

4.3.1 PERFORMANCE OF DIFFERENT BACKGROUND MODULES

As described in **Subsection 4.2.4**, I will show the impact of the choice of the background module on overall model performance. For that, I trained background modules with 6, 8, 10 and 12 states on the previously defined background training sets, and then either used them unchanged (termed *Standard*), or after the removal of states resembling the emission patterns of enhancers and promoters (*Reduced*). I determined the removed states in the *Reduced* models qualitatively by manually evaluating emission patterns. In addition, I derived a background module from the learned foreground modules using the same emission probabilities, but unitized transition probabilities (*FGtoBG*). With that, I want to investigate whether the foreground modules rely more on the different emission patterns compared to background genomic regions, or if their predictive power is due to the emphasis on the molecular structure, which is reflected in the constrained transition probabilities.

First, I evaluated the impact of the background modules on the overall predictive performance of the models using cross-validation on the mouse **ESC** data. With **AUPRCs** of 0.96 - 0.97, the *Standard* model clearly outperforms the alternative models (*Reduced*: 0.48 - 0.66, *FGtoBG*: 0.23, **Figure 4.9**). Further, the *Standard* model is robust to the number of states whereas the *Reduced* model performs better with a higher number of states.

Second, I assessed the effects of the background on the total number of predicted enhancers in a genome-wide matter (**Figure 4.10**). The *Standard* model predicts roughly between 4000 - 5000 enhancers and 8000 - 10,000 promoters. Both the *Reduced* and *FGtoBG*, however, predicted about ten times more enhancers, while detecting 17-64% less promoters compared to the equivalent *Standard* models in terms of the number of states.

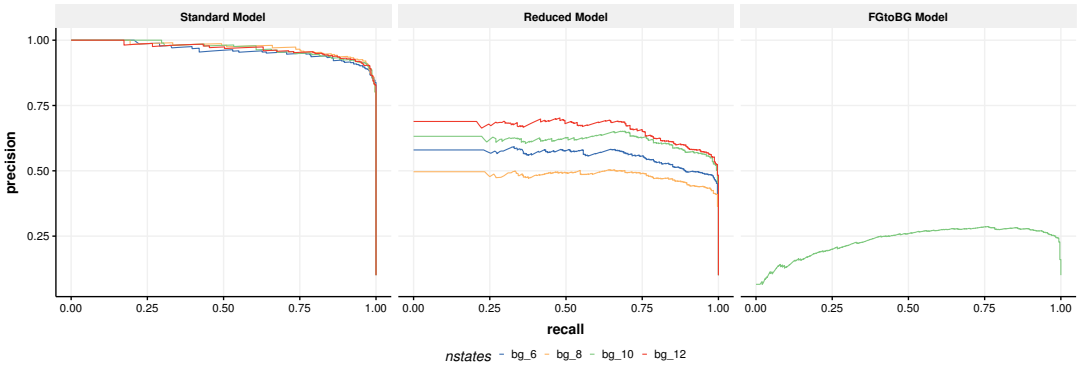

Figure 4.9: Precision recall curve using the *Standard* and alternative background modules with different numbers of states.

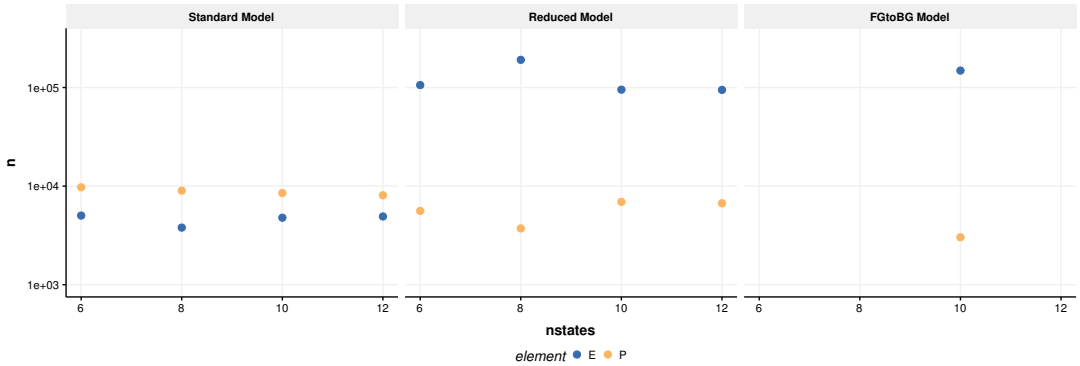

Figure 4.10: Number of predicted enhancers and promoters using the *Standard* and alternative background modules with different numbers of states.

Third, I investigated the enrichment of genomic signals over the predicted regions (**Figure 4.11**). By definition, there is only one implementation of the *FGtoBG* model with 10 states. Hence, I only compare the models with 10 background states. Moreover, due to the exceedingly high number of enhancer predictions in some models, I randomly sampled 3000 predictions from every model. The predicted promoters look very similar across all background modules and consistent with what we would expect from literature, i.e. high levels of **ATAC-seq** in the center flanked by high levels of **H3K27ac** and **H3K4me3**. The predicted enhancers, however, look substantially different between the different background modules, and only the *Standard* model predicts enhancers that are consistent with the expected properties, i.e. high levels of **ATAC-seq** in the center flanked by high levels of **H3K27ac** and **H3K4me1**. The high numbers of predicted enhancers using alternative background modules thus most likely suffer from a high false-positive rate. Thus, the enhancer- and promoter-like states in the *Standard* background that are not present in the alternative background modules likely contribute to an increased prediction specificity by competition with the foreground states.

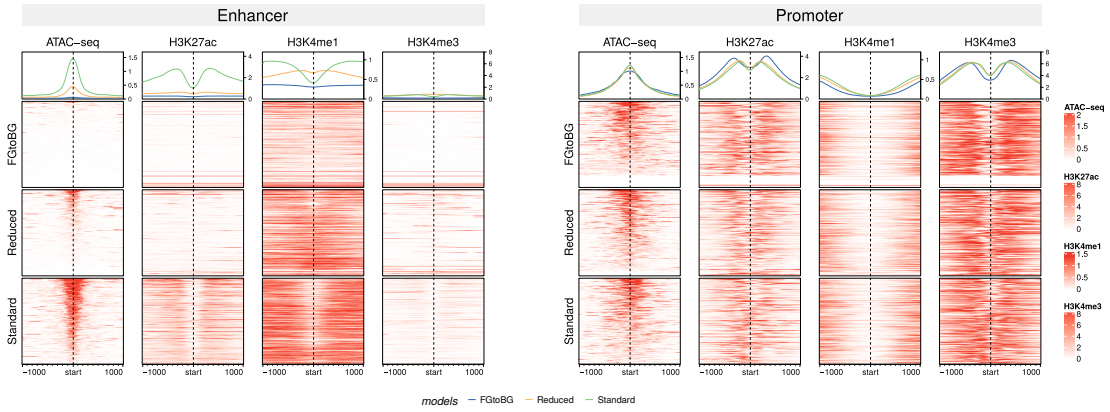

Figure 4.11: Feature heatmaps of predicted enhancers (left panel) and promoters (right panel) using different background modules. Average distributions are shown in the top sub-panels.

Combined, these findings substantiate the importance of choosing the right background module. While the sheer numbers of genome-wide predictions with the alternative models might look reasonable, their performance in cross-validation tests as well as the enrichment of genomic signals over the predicted regions clearly do not. In the next Subsection I will assess the method's performance in different settings. For that, I will use the *Standard* background module with 10 states.

4.3.2 PERFORMANCE WITHIN AND ACROSS SAMPLES

Once settled on the type of background module and after constructing the full model, it is of interest to quantify how well the model is able to predict enhancers. I describe the testing procedure in **Subsection 4.2.8**. In this Subsection, I will present the method's performance within and across samples and data types.

WITHIN SAMPLE VALIDATION **eHMM** is able to recall a high fraction of the labeled enhancers without capturing a lot of false positives. i.e. being very precise at the same time. This is demonstrated in **Figure 4.12** (upper panel), where blue lines depict **eHMM**'s precision and recall using variable prediction thresholds in cross-validations within samples. Notably, even low threshold values yield high precision while still capturing most enhancers from the test set. The good performance is quantified by the sample-specific **AUPRC** ranging between 0.947 and 0.971 (**Figure 4.14**).

CROSS SAMPLE VALIDATION Often, enhancer predictions are desired in specific samples for which it is infeasible to define a training set, e.g. because there is no

labeled data set available online for a particular cell type or developmental stage and executing the necessary experiments oneself is not practicable. Such scenarios require the option to train the method on one sample and apply it to another. I tested **eHMM**'s performance in cross-sample validation settings where I used the model trained on data from FANTOM5 in mouse **ESC** to predict FANTOM5 enhancers in mouse liver E12.5 and lung E16.5. I used quantile normalization (see **Subsection 4.2.10**) to account for potentially different read count scales between samples. As expected, method performance decreases slightly in across-sample validation compared to using a model trained on data from the same sample (**Figure 4.12**, lower panel). Areas under the precision-recall curve of 0.928 and 0.865 for liver E12.5 and lung E16.5, respectively, still show very satisfying results (**Figure 4.14**, blue dots). This demonstrates the method's great applicability with pre-trained models. Moreover, I show the suitability of the quantile normalization approach by comparing cross-sample validations with and without normalization. Quantile normalization helps to improve prediction quality with an increase in area under the precision-recall of 0.041 and 0.025 in liver E12.5 and lung E16.5, respectively.

eHMM does not require the user to set an arbitrary prediction threshold, but rather makes use of the *Viterbi decoding* algorithm. **Figure 4.12** displays shapes that illustrate the performance of *Viterbi decoding*. In both within and cross sample validations, *Viterbi decoding* yields precision and recall values residing in the top right corner of where the lines of variable prediction thresholds run, endorsing its application.

4.3.3 PREDICTION ROBUSTNESS AGAINST VARIABLE DATA SOURCES

To date, there is no gold standard enhancer set for any cell type. However, training and testing on data from one source only could potentially result in overfitting. To that end, I performed within and cross sample validation using an independent test set from EnhancerAtlas [212]. EnhancerAtlas integrates data from different sources and multiple enhancer signals, e.g. **DHS**, the enhancer associated histone acetyltransferase EP300 [222] or combined tracks such as **TFBSs** consisting of multiple motifs or the single or combined presence of histone modifications. The fundamental assumption of the EnhancerAtlas approach is that predictions from good quality data should coincide whereas those from low quality data should not. With that, EnhancerAtlas integrates numerous data sources and simultaneously weighs their individual impact according to their agreement with other sources.

I compared **eHMM**'s performance on regions from the EnhancerAtlas for the mouse samples **ESC** E14, liver E14.5 and lung E14.5 by measuring precision and recall (**Fig-**

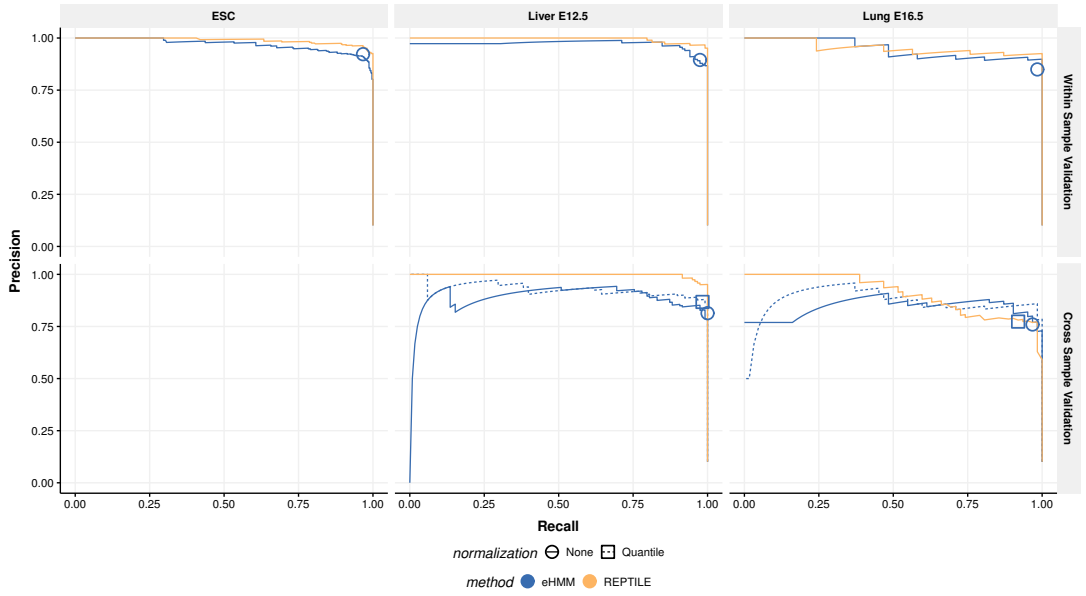

Figure 4.12: Precision recall curves for the supervised methods **eHMM** and REPTILE validated within and across samples on data from FANTOM5 in mouse **ESC**, liver E12.5 and lung E16.5. Shapes indicate prediction performance of the *Viterbi* algorithm. Lines represent precision and recall on posterior probabilities obtained from the *forward-backward* algorithm.

ures 4.13 and **4.14**, yellow dots). Compared to training and testing on data from the same source, **eHMM**'s performance dropped considerably in mouse **ESC** (**AUPRC** = 0.75). However, high performance was maintained in liver E14.5 (**AUPRC** = 0.93) and lung E14.5 (**AUPRC** = 0.86). Of course, the quality of the EnhancerAtlas data sets might vary itself, and thus it is necessary that these performance measures are put into perspective. I will do that in the next Subsection where I will benchmark **eHMM** in comparison with existing methods.

4.3.4 BENCHMARKING

Identifying regulatory elements has been a central objective in computational biology for decades, and numerous software packages exist that tackle this task relying on various experimental data [196, 197]. In this Subsection I compare the prediction performance of **eHMM** to a selection of existing methods, i.e. ChromHMM [223], EpiCSeg [211] and REPTILE [224]. I chose these methods for a variety of reasons. First, ChromHMM is a well-established and widely used method that learns a hidden Markov model based on binarized input data in an unsupervised fashion. **EpiCSeg** presents another unsupervised HMM that also provided the foundation of the implementation of **eHMM**. In contrast to ChromHMM, it models the read count data

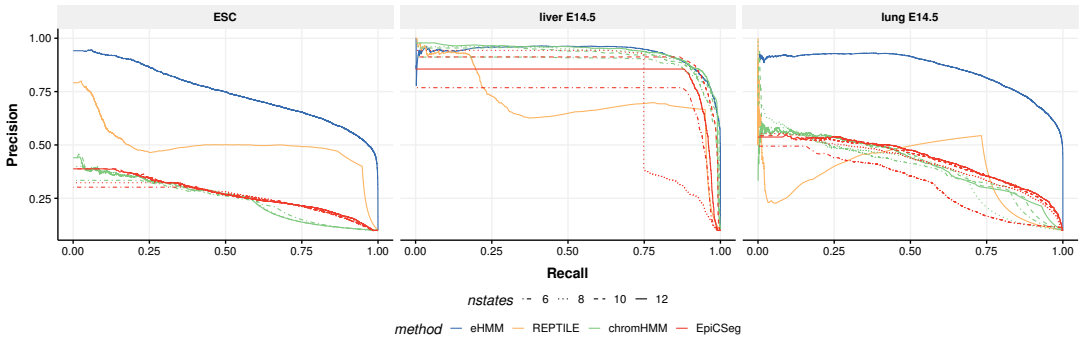

Figure 4.13: Precision recall curves for all tested methods validated on data from the EnhancerAtlas in mouse **ESC**, liver E14.5 and lung E14.5. Lines represent precision and recall on posterior probabilities obtained from the *forward-backward* algorithm.

using a negative multinomial distribution instead of binarized data. Together, these two methods allow the comparison of my supervised **HMM** to two unsupervised **HMMs** and thus to investigate the benefit of supervision. Finally, REPTILE is a supervised method using a random forest classifier, which I train with the same training data as **eHMM** in order to study the differences between two supervised methods that use a different mathematical model. As shown in the article by He et al. [224], REPTILE is state-of-the-art as it outperforms previous methods. It therefore certainly serves as a challenging competitor to **eHMM**. Unfortunately, I was not able to test ChroModule (see **Section 4.1**) as the authors do not provide the software.

As ChromHMM and **EpiCSeg** are unsupervised methods, I applied them to whole genome data with different numbers of states (6, 8, 10 and 12). With the resulting parameters I computed the maximum posterior probability of every state in the test regions and report only the best performing state. Of course, state selection based on performance on a test set is not within the usual application of these methods. Typically, states are selected by eye, potentially introducing an additional source of error and bias. However, for the sake of benchmarking I assume optimal state selection. I tested REPTILE and **eHMM** within cell types using 5-fold cross-validations on FANTOM5 data and across cell types and data sources by validating the performance of a model trained on aforementioned mouse **ESC** training data on test regions from FANTOM5 and EnhancerAtlas.

eHMM performs equally well or better than all other methods in all but one scenarios (**Figures 4.12** to **4.14**). Only in the cross sample validation in liver, REPTILE performs slightly better than **eHMM**. Generally, the supervised methods **eHMM** and REPTILE tend to outperform the unsupervised ChromHMM and **EpiCSeg** in most test settings, demonstrating the benefit of supervised learning for the task at hand.

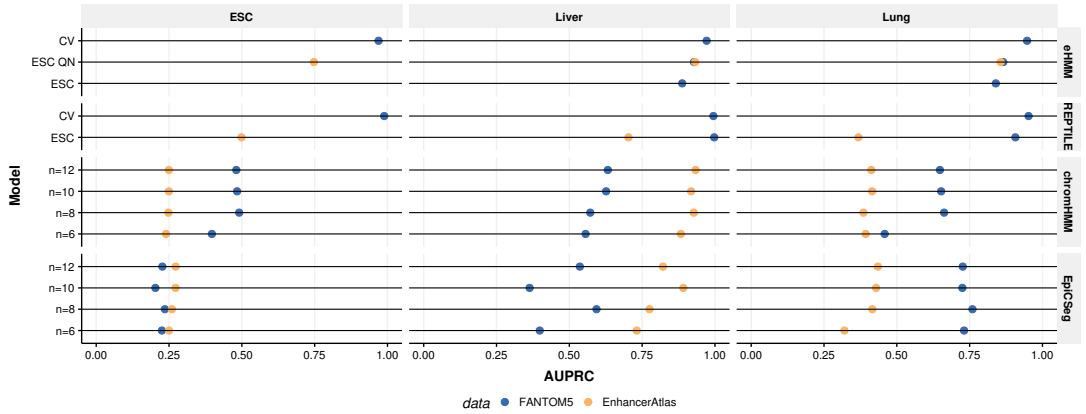

Figure 4.14: AUPRC for all tested methods validated on data from FANTOM5 (blue) and EnhancerAtlas (orange). Note that the columns 'Liver' and 'Lung' include data from different developmental stages, i.e. E12.5 and E14.5 for FANTOM5 and EnhancerAtlas, respectively, for 'Liver', and E16.5 and E14.5 for FANTOM5 and EnhancerAtlas, respectively, for 'Lung', listed in detail in **Appendix A.2.1**. Legend acronyms: **CV** - within-sample 5-fold cross-validation. **ESC QN** - cross-sample validation using a model trained on ESC data including quantile normalization. **ESC** - cross-sample validation using a model trained on ESC data without normalization. **n** - number of states in HMM.

Benchmarking aims to report the methods' performances in different test scenarios. As indicated in previous Subsections, this includes the performance within and across samples and data sources. Overall, **eHMM** proves to be very robust against varying test scenarios as performance measures are more or less maintained when altering the sample, the data source or both. In contrast, REPTILE generally performs well when the data source for training and testing is the same, but struggles when it is varied (**Figure 4.14**), suggesting overfitting of the learned models on the FANTOM5 data. The tested unsupervised models depend strongly on the data quality and less on the method (ChromHMM or **EpiCSeg**) or the chosen hyperparameter (the number of states). For example, all tested unsupervised models predict with **AUPRC** < 0.5 on EnhancerAtlas lung E14.5 data, but with **AUPRC** > 0.75 on EnhancerAtlas liver E14.5 data. Together, these results show underline the robustness of **eHMM** under different types of validation setups.

4.3.5 WHOLE GENOME PREDICTIONS IN MOUSE EMBRYONIC STEM CELLS

The testing procedures described so far allowed the quantitative measurement of the method's ability to fulfil its purpose - to identify enhancers. Naturally, the next step is to apply the method to a full genome data set to identify enhancers genome-wide. **eHMM** predicts 5357 enhancers and 8040 promoters in mouse **ESC**. Depending on the chosen prediction threshold $c \in [0, 1]$, REPTILE predicts between 2604 ($c = 0.9$)

and 12,830 ($c = 0.1$) enhancers. With a varying number of states n, ChromHMM finds between 19,643 ($n = 12$) and 88,716 ($n = 6$) enhancers, EpiCSeg between 37,911 ($n = 12$) and 103,293 ($n = 6$). In this Subsection I will discuss the properties of **eHMM**'s identified enhancers and promoters in mouse **ESC** as depicted in **Figure 4.15**.

ATAC-SEQ AND HISTONE MODIFICATIONS The identified regulatory regions exhibit the anticipated average relative enrichment or depletion of the features used for training, i.e. **ATAC-seq**, **H3K27ac**, **H3K4me1** and **H3K4me3**. Both predicted enhancers and promoters show a bimodal enrichment of the respective **histone modifications** with a central unimodal enrichment of **ATAC-seq**. Predicted enhancers and promoters show high and low ratios of **H3K4me1/3**, respectively. These observations provide evidence for the initial biological assumptions that motivated and shaped the model's architecture.

BINDING OF TRANSCRIPTION FACTORS AND CHROMATIN REMODELERS Predicted enhancers show enriched binding of **ESC**-specific transcription factors Nanog, Oct4 and Sox2. This effect is present, but less pronounced in promoters and thus in line with the hypothesis that enhancers are more lineage-specific than promoters, and that promoters can be regulated by different sets of lineage-specific enhancers depending on the cell type [225].

In addition, predicted enhancers show elevated levels of the **HAT** p300, an enzyme involved in transcriptional regulation via chromatin remodeling. p300 was reported to deposit **H3K27ac** [49] mainly at **TSS**-distal sites [53] and its binding locations are associated with active enhancers [69]. Interestingly, although promoters exhibit higher levels of **H3K27ac** on average, p300 indeed seems to bind more exclusively to enhancers, suggesting that other **HATs** might be responsible for the deposition of the bulk of **H3K27ac** at promoters.

Binding events of **CCCTC-binding factor (CTCF)**, a protein involved in the regulation of the three dimensional chromatin structure [226] and often co-occurring with the borders of **TADs** [16], are enriched in enhancers, implying the enhancers' role in the mediation of enhancer-promoter contacts and **DNA** looping [227, 228]. The depletion of **CTCF** binding in predicted promoters might hint at the incompleteness of the set. Some promoters show very asymmetrical features, e.g. because they are highly one-directional. **eHMM** assumes symmetry in enhancers and promoters and might misclassify such cases. Promoters at **TAD** boundaries, however, are expected to be

one-directional by nature. However, **eHMM**'s main purpose is to identify enhancers, not promoters, which is why I only acknowledge this shortcoming of the promoter module, but do not tackle it.

DNA METHYLATION AND SEQUENCE CONSERVATION Both enhancers and promoters show a dip in **DNA** methylation measured by **MeDIP-seq**. This effect appears to be stronger in predicted promoters, confirming previous studies that suggest that **DNA** methylation levels negatively correlate with **H3K4me3** [229] and that high-CpG-density promoters generally contain a core region of unmethylated CpGs [230]. Predicted promoters exhibit increased sequence conservation across species as depicted by phastCons conservation scores [218]. Enhancers indicate this feature as well, but to a much lower extent, agreeing with Villar et al. [97] who showed that enhancers evolve much more rapidly than promoters.

RNA POLYMERASE II Finally, promoters exhibit high bimodal levels of **RNAP II**, indicating transcription initiation events in both directions. Enhancer elements show a similar pattern but at lower levels, confirming that the input data from FANTOM5 reflects the information about the bidirectional transcription initiation which had originally motivated our choice of the training set.

Peaks of features that are enriched unimodally are especially sharp in enhancers, indicating that enhancer predictions are centered well on the true accessible chromatin. In the next Subsection, I will quantify this spatial accuracy of **eHMM**'s enhancer predictions.

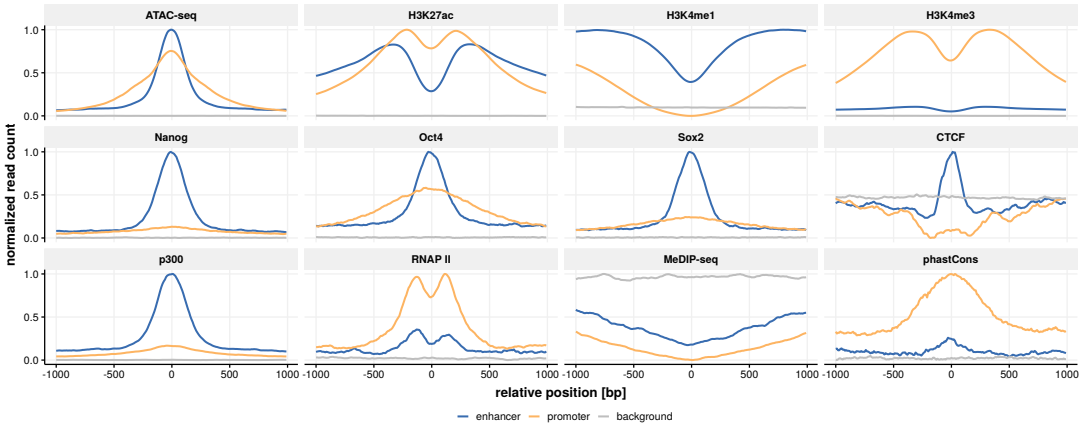

Figure 4.15: Mean feature distributions of genome-wide predicted enhancers and promoters in mouse **ESC**.

4.3.6 SPATIAL ACCURACY

In previous Subsections I have assessed **eHMM**'s capability to predict enhancers in quantitative validations on predefined test sets as well as a qualitative analysis of the features of genome-wide predictions. Researchers that are interested in predicting enhancers prior to downstream analyses might also care for the resolution of the predictions, i.e. the spatial accuracy of the predicted enhancer locations. Active enhancers are expected to contain accessible chromatin, which is where **TFs** bind and establish the enhancer's function. Chromatin accessibility can be measured by **ATAC-seq**, and the distance of an enhancer prediction to the nearest **ATAC-seq** peak thus provides a measure of spatial accuracy. **eHMM** predictions are on average around seven times closer to the center of an accessible region compared to REPTILE (median of 42 **bps** and 343 **bps**, respectively, **Figure 4.16**). Including other features such as **DNA** methylation might improve REPTILE's spatial prediction accuracy, however, at the expense of requiring additional data. Moreover, even the most complex REPTILE models do not achieve average distances lower than 111 **bps** and 65 **bps** in mouse embryonic tissues E11.5 and human H1 derived cells, respectively [224]. This highlights **eHMM**'s outstanding resolution which is accomplished by the model architecture, namely the distinction between nucleosomal and accessible states.

4.3.7 PREDICTED ENHANCERS ARE TSS-DISTAL

Promoters and enhancers are mainly distinguished by the degree of methylation of **H3K4**. These residues are generally trimethylated in the immediate proximity of a promoter's center. When moving away from a promoter's center, the likelihood of **H3K4** being trimethylated drops fast and is replaced by monomethylation. At that point, nucleosomes resemble those of a typical enhancer. However, these nucleosomes are in the periphery of promoters and do not border accessible chromatin themselves. **Figure 4.17** illustrates this problem, showing an example gene where **eHMM** correctly predicts a promoter at the upstream end of a transcribed gene, while REPTILE misclassifies the adjacent region as an enhancer. I quantified this effect by calculating the distances of genome-wide predicted enhancers to the closest annotated TSS. Distances of predicted enhancers to the closest annotated TSS are unimodally distributed in the case of **eHMM** with an interquartile range spanning from 11 kb to 85 kb (**Figure 4.16**). Enhancers predicted by REPTILE exhibit an additional mode that centers at approximately 1 kb and most likely represents false enhancer predictions adjacent to promoters.

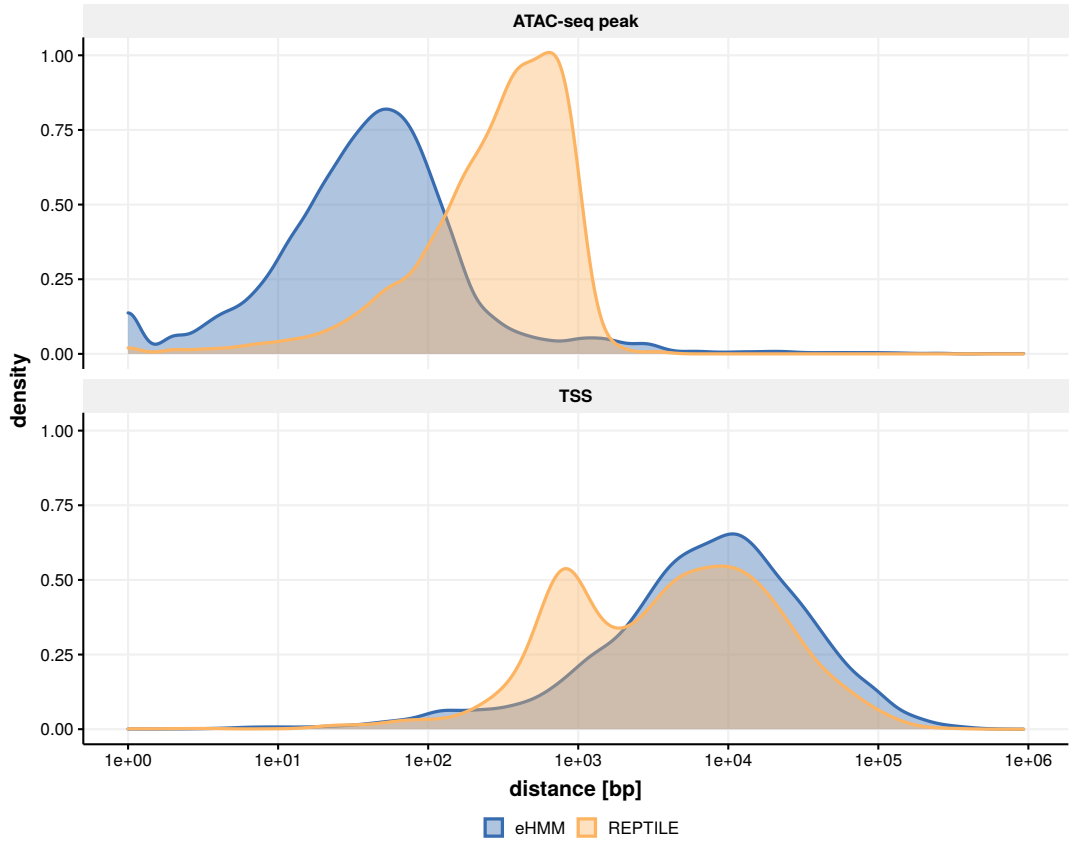

Figure 4.16: Distance distributions of predicted enhancers to closest **ATAC-seq** peak (MACS2) and TSS (UCSC knownGene database) in mouse **ESC** for **eHMM** and REPTILE (threshold = 0.9).

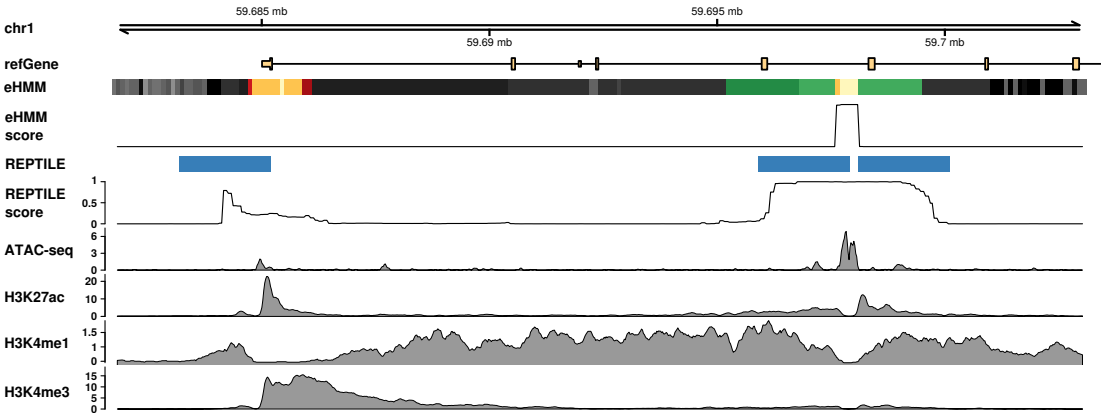

Figure 4.17: Example genomic region with predictions from **eHMM** and REPTILE (threshold = 0.5). The color code in the **eHMM** segmentation track is equal to **Figure 4.4**, i.e. green: enhancer nucleosome, red: promoter nucleosome, yellow: chromatin accessibility, gray: background.

4.3.8 RUN TIME

I estimated empirical run times for model training and prediction on mouse **ESC** data and compared them to those of REPTILE, **EpiCSeg** and ChromHMM. All methods ran on 21 cores in parallel (one for every chromosome) as far as the respective implementation allowed it. Run times per core are shown in **Table 1**. REPTILE uses the least total CPU time, but the longest real time, indicating a lack of efficiency in leveraging multithreading. **eHMM** completes training and prediction in less than 47 seconds (real time), which is similar to **EpiCSeg** (36 seconds) and ChromHMM (50 seconds). REPTILE takes almost twice the amount of time (90 seconds).

Table 1: Run times

	Real time [s]			CPU time [s]		
Method	Training	Prediction	Total	Training	Prediction	Total
eHMM	2.961	43.636	46.597	15.337	155.820	171.157
REPTILE	1.461	89.456	90.917	5.162	140.388	145.550
EpiCSeg			36.327			352.294
ChromHMM			50.401			282.909

4.4 DISCUSSION

I developed **eHMM** with the goal of detecting active enhancers with variable lengths throughout mammalian genomes. **eHMM** features three modules for enhancer, promoter and background prediction, each being trained in a supervised fashion on predefined training sets. The enhancer and promoter modules consist of a particular architecture that captures the biological topology of these regulatory elements, i.e. a central accessible stretch of **DNA** flanked by nucleosomes to each side. The method performs well in cross-validation tests, showing that the proposed physical model is present in the data and captured by **eHMM**. Moreover, **eHMM** incorporates a quantile normalization step that makes it well applicable across samples, e.g. a model trained on one cell type or developmental stage can be used for predictions on another. Based solely on the area under the precision-recall curve as a performance measure, **eHMM** achieves similar results as the top-performing state-of-the-art software REPTILE when testing on the FANTOM5 data set, and outperforms it when validating on regions from the EnhancerAtlas. These results suggest overfitting of the models learned by REPTILE and underline the robustness of **eHMM**'s predictions over different validation setups. Notably, there are apparent performance differences

between cell types. In particular the prediction performance on **ESC** is generally lower compared to lung and liver. This is likely due to the fact that EnhancerAtlas regions were predicted on the basis of agreement of different source tracks such as **TFBSs**, **eRNA**, **histone modifications**, chromatin accessibility and more. Here, I use only chromatin accessibility and **histone modifications**, and I would thus expect the tested methods to perform best in cell types where these features were most informative for the EnhancerAtlas predictions. The results suggest that **ESC** regions in the EnhancerAtlas were not primarily predicted on the basis of the features used in this study. The outcome of unsupervised methods such as ChromHMM and **EpiCSeg** is uncertain as they perform well in some conditions and poorly in others, and it is not apparent how to judge the quality of a segmentation without a test set. In addition, state interpretation is not trivial and highly affects the prediction quality.

eHMM's genome-wide enhancer and promoter predictions in mouse **ESC** exhibit expected properties, confirming prediction quality on a whole-genome level. For example, lineage-specific **transcription factors** are enriched at enhancers, and promoters exhibit low **DNA** methylation levels and an abundance of **RNAP II**. In contrast to previous work focusing on sequence conservation in cis-regulatory regions [176, 193–195, 231], these results show that the sequence of predicted enhancers is less likely to be conserved in comparison to predicted promoters. This seeming contradiction between observing strong binding of lineage-specific transcription factors and low levels of sequence conservation could suggest functional conservation while the enhancers' genomic locations are highly dynamic in evolutionary terms as suggested by Schmidt et al. [232] and others [98, 99, 101], manifesting itself in a lower sequence conservation across species.

The lower number of predicted enhancers with the supervised methods **eHMM** and REPTILE reflects their higher specificity compared to the unsupervised methods ChromHMM and **EpiCSeg**. While REPTILE enforces this specificity rather arbitrarily by calling only the most certain enhancer among multiple neighboring predictions, **eHMM** achieves this by the potential presence of enhancer- and promoter-like states in the background module that compete with the topology-respecting foreground module. **eHMM** thus ultimately reduces the false-positive rate by emphasizing the importance of the enhancers' molecular structure, which in turn results in higher spatial accuracy. Further, **eHMM** returns the most likely path according to the *Viterbi decoding* algorithm and therefore does not require the definition of an arbitrary prediction threshold.

Many methods often predict enhancers right next to promoters where the promoter-specific histone modification **H3K4me3** decreases while **H3K4me1** re-emerges. The

implemented promoter module as well as the aforementioned model topology enables **eHMM** to distinguish between the two regulatory elements and to refrain from calling enhancers in promoter-associated regions merely on the basis of a decreasing promoter signal. In addition, **eHMM** provides a high resolution of predicted regions, allowing the accurate targeting of regulatory subunits such as nucleosomal or accessible regions for potential downstream analyses.

Moreover, **eHMM** allows inspection of model parameters that provide information about both transition dynamics between states and each state's signal emission distribution, standing in contrast to *black box* methods. These properties facilitate interpretability of the learned parameters and the predicted regions, and allow us to draw biological conclusions. Finally, I showed how to use hidden Markov models in a supervised fashion with functional genomic data, and how different models learned on various training sets can be combined in order to obtain one global model containing supervised modules with well-defined topologies.

Taken together, the minimal feature requirements, good performance within and across samples, the predictions' high spatial accuracy as well as interpretability and resolution make **eHMM** a very powerful and feasible tool for enhancer prediction. However, while the method tackles problems and shortcomings of earlier methods, it comes with certain drawbacks itself. Optimizing the performance of a computational method is often a trade-off between various objectives. For example, **eHMM**'s high confidence predictions come at the price of low sensitivity. Downstream analyses relying on computationally identified enhancers may depend on high sample sizes, and in those cases **eHMM** may not be the method of choice. This is particularly the case when performing statistical analyses for single loci, e.g. motif analysis of all enhancers in a **TAD** as in **Subsection 5.3.4**. Moreover, **eHMM** requires minimal input data, a property that is especially useful when data availability is scarce. However, other methods may be more suitable in case additional data such as **DNA** methylation is available. Extending **eHMM** to optionally integrate additional data presents an opportunity to expand the applicability of the method. Finally, **eHMM** is designed for bulk functional genomics data and is not suitable for handling sparse read count data from single cell experiments such as **scATAC-seq**. It will certainly be interesting to conceptualize a method using single cell data together with maintaining the advantages achieved in this work in the future and thereby addressing the identification of cell type-specific enhancers.

5 FUNCTIONAL CONSERVATION OF CIS-REGULATORY ELEMENTS

5.1 MOTIVATION

IDENTIFYING FUNCTIONALLY CONSERVED ELEMENTS Two species share parts of their genomes that have been conserved since their evolutionary paths diverged. In a simplified view, the more time has passed since the common ancestor and the faster the adaptation or the higher the mutation rate, the less frequent are those conserved elements. Conserved elements are typically functional and their conservation the very consequence thereof. In **Subsection 2.2.5** I introduced the topic of sequence conservation and noted that enhancers display very heterogeneous levels thereof [97]. The results presented in **Chapter 4** confirmed these previous findings by showing diverse levels of sequence conservation in predicted enhancers (**Subsection 4.3.5**). Moreover, there have been individual reports of enhancers that are not conserved in sequence but rather in function [86, 95, 96, 98, 99, 101]. Measured by the multitude of such findings, functional conservation in absence of sequence conservation seems to be a widespread phenomenon.

In 2011, Taher et al. [100] presented an attempt to systematically identify functionally conserved elements. They did so by exploiting the fact that for any given genomic location, mutation rates and selective pressure may vary between species. This means that even if the sequence in putative functional orthologs between two species diverged to an extent that renders them unalignable, both of these orthologs might be alignable to a third species. In that case, the sequence of the third species serves as an "orthology tunnel" between the originally compared species. I will use the term *bridging species* for such a third species harboring an orthology tunnel. Using this approach with the western clawed frog (*Xenopus tropicalis*) as their bridging species, they found approximately 300 pairs of functional orthologs with diverged sequences between human and zebrafish (*Danio rerio*) that are likely to share common ancestry and regulatory activity based on **TFBS** composition. The very fundamental factor for their approach is that different genomic locations vary in mutation rates and selective pressures between species. This entails that a fixed bridging species will only reveal a subset of all true functional orthologs because many of them might be tunneled through a different bridging species. Optimizing the choice of the bridging species and, moreover, using multiple bridging species in sequence thus may enhance the capability of identifying functional orthologs beyond sequence conservation.

In this Chapter I will present two methods that optimize the compilation of bridging species for a given genomic location under slightly different aspects (**Subsections 5.2.1** and **5.2.2**).

PROJECTING GENOMIC COORDINATES OF NON-ALIGNABLE ELEMENTS For the task at hand the goal is to achieve accurate projections of unalignable genomic regions between a *reference* and a *target* species. Alignable and non-alignable regions are interspersed, i.e. non-alignable regions are generally located between alignable regions except for genomic regions at the chromosomal margins. Two methods presented in **Subsections 5.2.1** and **5.2.2** make use of that fact and assume that if a non-alignable genomic element is located between two alignable sequences in a syntenic neighborhood in the reference species, then it will also be located between these two alignable regions in the target species if it exists. In other words, non-alignable but functional sequences within a **GRB** are either lineage-specific or they are functionally conserved and thus also expected to maintain synteny, i.e. the co-localization and conservation of order between the two species. For this assumption to hold true, the flanking alignable regions, which I term *anchor points*, are required to exhibit shared synteny, too. In that case, the anchor points map to roughly the same genomic neighborhood in the target species. Especially **CNEs** clustering in **GRBs** have been found to be largely syntenic and thus collinear [233].

The linear distance between any two anchor points depends on the similarity of the compared genomic regions which is a function of the evolutionary distance between the two species and their respective local mutation rates. For example, let us imagine a random genomic location in mouse that is neither alignable to human nor to zebrafish. On average, its anchor points to human are expected to be closer than its anchor points to zebrafish, given that human and mouse diverged roughly 90 **million years ago (mya)** and share much more genetic information than mouse and zebrafish who diverged ~435 **mya** [234]. This example is visualized in **Figure 5.1 A**. The distribution of **CNEs** in **GRBs** as depicted in **Figure 5.1 C** shows a distinct pattern that changes with the evolutionary distance of the compared species (**Figure 5.1 B**). Most distributions are bimodal with the first mode at short distances in the order of 100 **bps**. This mode is consistent in all compared species and likely represents tight clusters of **CNEs** that potentially form subdomains within a larger unit, the **GRB**. The second mode only exists with larger evolutionary distances and its position positively correlates with these very distances, representing the sparser distribution of conserved elements between species with high evolutionary distances and thus their larger spacing. This is summarized in **Figure 5.1 D** showing an exponential relationship between the divergence time of two species and the average distance between

two **CNEs**. Notably, elephant shark (*Callorhinchus milii*) stands out with lower **CNE** distances than zebrafish. This is owed to the fact that elephant shark has the slowest evolving genome among all examined vertebrates and thus resembles the last common ancestor more than zebrafish does [235]. This highlights that not only time but also individual mutation rates affect anchor point distributions.

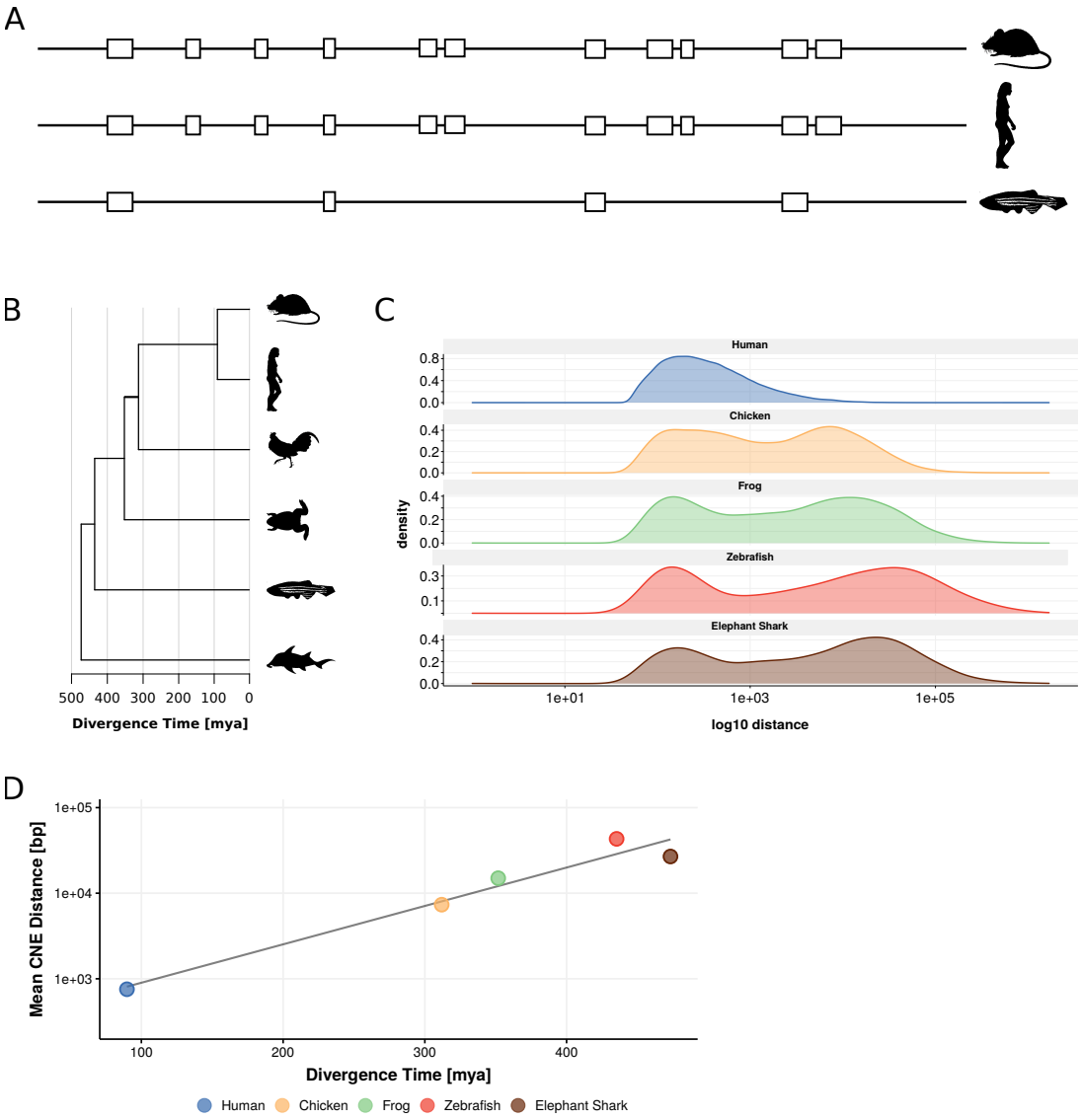

Figure 5.1: Distribution of **CNEs** between two species in relation to their evolutionary distance. **A** Schematic illustration of **CNE** densities between mouse, human and zebrafish. **B** Phylogenetic tree showing evolutionary distances between mouse and five species (human, chicken, frog, zebrafish and elephant shark). Evolutionary divergence times according to Kumar et al. [234]. **C** Distributions of inter-**CNE** distances in comparisons from mouse to five species. **D** Relationship between divergence time and the average distance between two **CNEs** shared between mouse and one of five species.

FUNCTIONAL CONSERVATION OF TOPOLOGICAL CHROMATIN STRUCTURES In **Section 2.3** I discussed the nature of highly conserved genomic elements, so called **CNEs**, and their clustering into **GRBs**. Many **CNEs** have been shown to act as enhancers on the regulation of key developmental genes [90]. Not only are **CNEs** conserved in their sequence but also constrained in their order and they are largely collinear [233]. This global constraint likely represents higher-order organization such as the three-dimensional chromatin structure. A recent study reported that **GRBs** colocalize with **TADs**, suggesting a functional role of **CNEs** in the regulation of the spatial chromatin structure [112]. **CNEs** identified in different lineages with large evolutionary distances often lack sequence conservation due to gradual and continual **CNE** turnover [236]. Regardless, **TAD** structures are often conserved even in **GRBs** with high **CNE** turnover, suggesting that the three-dimensional organization plays an important role and can be maintained even in the absence of sequence conservation [237, 238]. I investigated this context by mapping epigenomic data representing higher order chromatin structure from mouse onto zebrafish genomic coordinates and assessed the conservation of topological structures on the epigenomic level beyond sequence conservation. In **Subsection 5.3.4** I will present the discovery of epigenomically conserved subdomains and further provide evidence for the functional equivalence of the regulatory elements within those subdomains.

The development of the limb bud in vertebrates is one of the most studied experimental frameworks for investigating the molecular and morphogenic processes during organogenesis [239]. In addition, it serves as a model for studying evolutionary innovations, for example the appearance of digits during the fin-to-limb transition in the evolution of terrestrial vertebrates [240, 241]. It is especially suitable for studying the effect of genetic manipulations in limb-specific loci as the consequences of those may lead to malformations but are typically not lethal, and thus provides the ideal framework for the field of comparative genomics. In **Chapter 4** I presented a method for identifying enhancers using functional genomics data. In **Subsection 5.3.1** I will describe how I apply this method on data from the developing limb in mouse and use the identified enhancers as subjects in an attempt to identify putative functional orthologs between mouse and chicken.

5.2 METHODS

Depending on the context it is either desirable to project a given genomic location as accurately as possible or to project multiple neighboring regions subject to the constraint of synteny. An example for the first is the projection of an unalignable enhancer from one species to another, e.g. for identifying functionally conserved or-

thologs in the absence of sequence conservation. An example for the latter is the projection of a whole **GRB**. I will discuss these two approaches under the terms **Independent Point Projection (IPP)** and **Syntenic Anchor Point Propagation (SAPP)** in the subsequent parts of this Section.

5.2.1 INDEPENDENT POINT PROJECTION (IPP)

In syntenic genomic neighborhoods, projecting a genomic location from a given species to another in the absence of sequence conservation can be achieved by interpolating its relative position between two alignable anchor points. Naturally, this quickly becomes inaccurate when the anchor points are far apart as often is the case when comparing distantly related species. Using a bridging species may increase the density of anchor points and thus improve projection accuracy. In fact, the closer a genomic position is located to an anchor point, the less uncertain a projection using that anchor point. Therefore, it is desirable to choose the bridging species individually for every genomic location so that the distance to one of the anchor points is minimized. **Figure 5.2 A** schematically illustrates the benefit of using a bridging species on the projection accuracy for an example projection from zebrafish to mouse.

Finding an optimal bridging species depends on the distance of the genomic location x to the anchor points. However, it is not sufficient to only consider the distances in the reference species. If x_R is projected from the reference species R to x_B in a bridging species B and subsequently from B to x_T in the target species T, then the projection uncertainty will propagate through the species path $R \rightarrow B \rightarrow T$. The final projection accuracy will thus depend on both the distances from x to its closest anchor point $a_{R,B}$ as well as from the intermediate projection x_B to its anchor point $a_{B,T}$. **Figure 5.2 B** illustrates a scenario with a bridging species that minimizes only the distance in the reference, but not in the bridging species. A sequence of multiple bridging species will thus propagate the uncertainty of every step through the path.

Finding the optimal set of bridging species presents a shortest path problem. In graph theory, the shortest path problem is the search for a path between two vertices of a graph such that the sum of the lengths of its constituent edges is minimized. I discuss the topic in **Section 3.2** and introduce **Dijkstra's Shortest Path Algorithm**. Here, the vertices are the species and the weighted edges between them represent the distance of a genomic location to its anchor point. A schematic visualization of such a graph for a set of 15 species is given in **Figure 5.3**. The closer a genomic location to an anchor point, the smaller the distance in the graph for this species pair. In the following paragraph I establish a distance scoring function designed for that purpose.

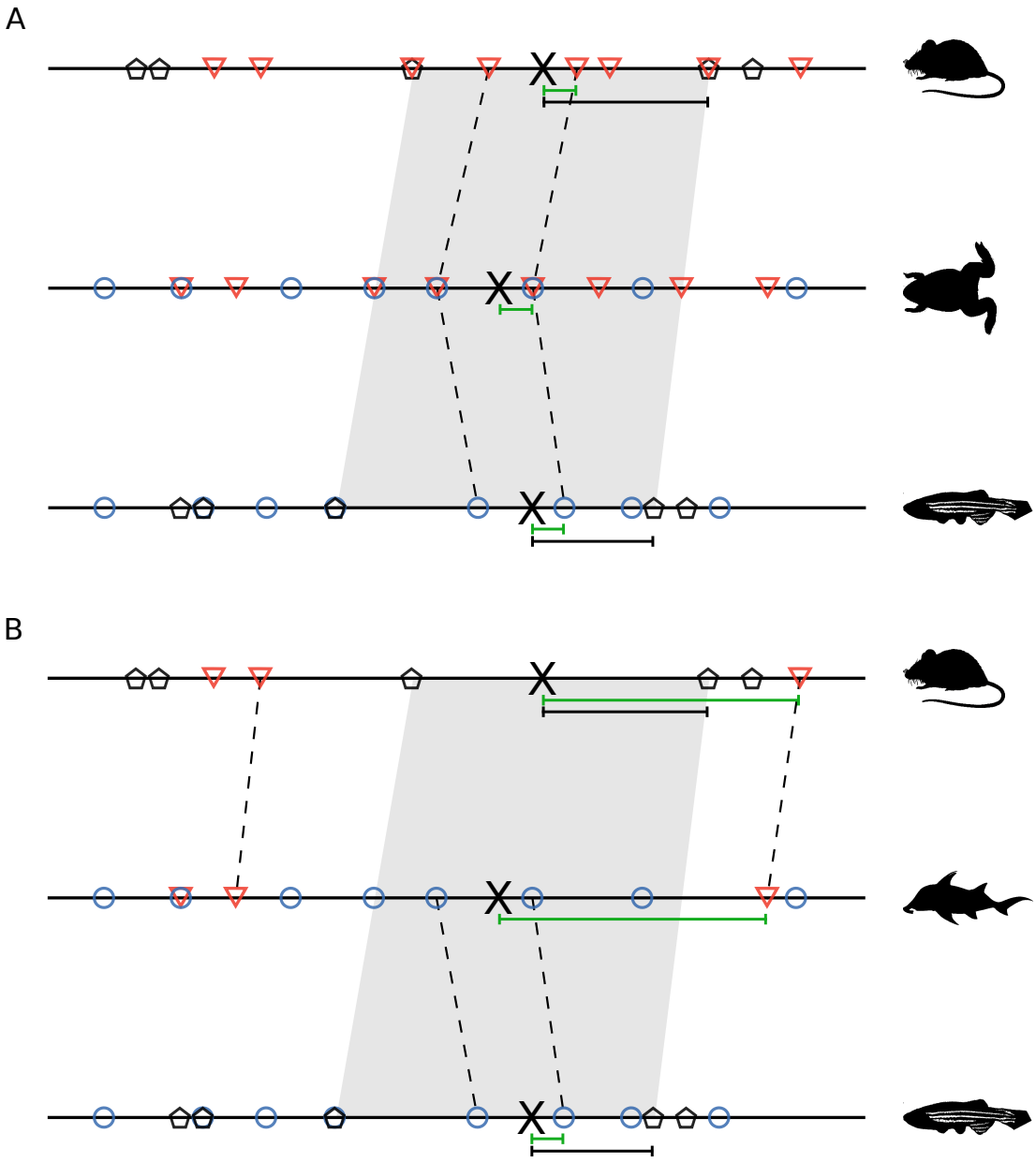

Figure 5.2: Schematic illustration of the **IPP** algorithm from zebrafish to mouse using different bridging species. Regions (X) in species other than the reference are found by linear interpolation between anchors. Symbols represent anchors between two species with their colors and shapes highlighting species pairs (e.g. black pentagons correspond to anchors between zebrafish and mouse). The grey shaded background depicts the span of the direct anchors, dashed lines mark the spans of the anchors through the bridging species. Horizontal bars represent distances from the closest anchor to the genomic coordinate and their colors distinguish direct projections (black) and projections via bridging species (green). The algorithm is outlined in **Algorithm 1**. **A** Bridging via frog minimizes distances to anchor points. **B** Bridging via elephant shark minimizes distance to anchor point in zebrafish, but not in elephant shark.

Independent point projections optimize projection accuracy for single independent genomic locations. However, in certain circumstances the genomic locations of interest are not independent and this method not suitable. I discuss the projection of dependent genomic locations in **Subsection 5.2.2**.

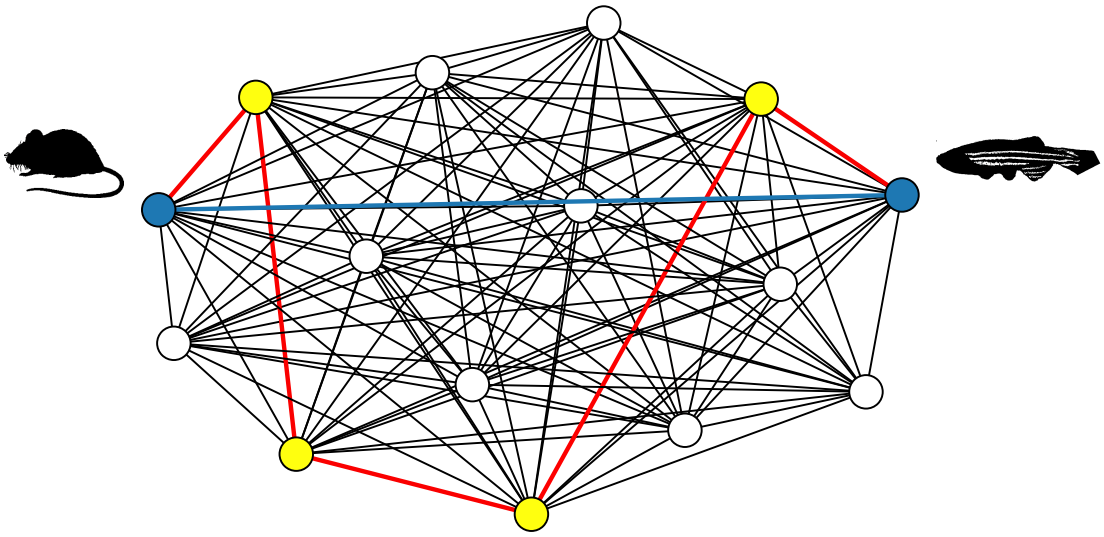

Figure 5.3: Species graph. Vertices represent different species and the edges are weighted by the distance scores between the vertices they connect. The highlighted paths mark two possible paths through the graph from mouse to zebrafish.

DISTANCE SCORING FUNCTION The **IPP** algorithm aims at maximizing projection accuracy by minimizing distances from the initial genomic location and all potential intermediate projections to their respective anchor points. For that, I established a scoring function that reflects those distances and returns values between 0 and 1. A score of 1 means that a genomic location x overlaps an anchor point a. The score decreases exponentially as the distance $|x - a|$ increases. For a single species comparison, the function is defined as follows:

$$f(x) = \exp\left(-\frac{\min(|x - a^{(1)}|, |x - a^{(2)}|)}{g\, s}\right), \tag{5.1}$$

with g denoting the genome size of the respective species and s a scaling factor. s can be determined by defining a *distance half life* d_h, i.e. the distance $|x - a|$ at which the scoring function is to return a value of 0.5. Solving $f(d_h) = 0.5$ for s we get

$$s = -\frac{d_h}{g\, \log(0.5)}. \tag{5.2}$$

Figure 5.4 shows the distance function for different distance half life values around two anchors that are located 200 **kbp** apart.

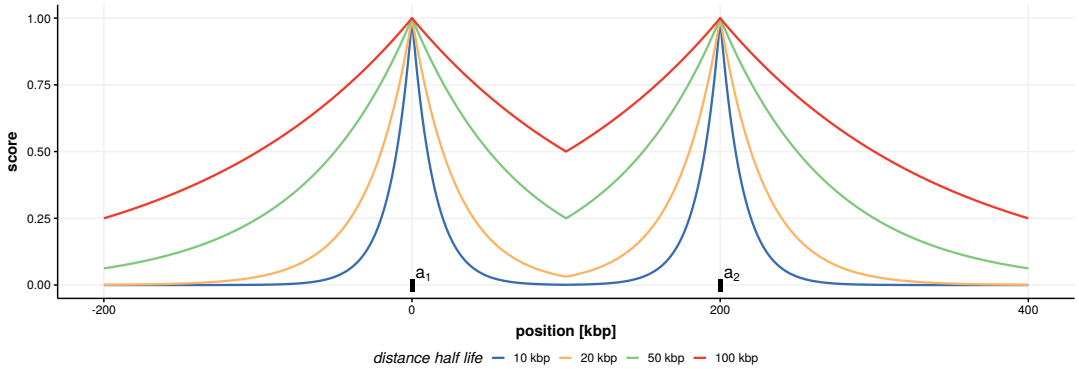

Figure 5.4: Course of the distance scoring function as defined in **Equation 5.1** with varying scaling factors around anchor points a_1 and a_2 located at 0 and 200 **kbp**, respectively. The scaling factors are based on distance half life values between 10 and 100 **kbp**. Shorter distance half lives let the score decrease faster when moving away from an anchor point.

It is important to note that when projecting from a reference to a target species through a path comprising a given set of bridging species, the scaling factor is only determined once for the reference species and kept constant for every subsequent bridging species in the path. That way, the genome size scaling factor g adjusts for different genome sizes in the respective bridging species.

The length l_p of a path p through the species graph G negatively correlates with the scores for each edge in p. Thus, l_p is obtained by multiplying the distance scoring function for every species in that path and subtracting that value from 1:

$$l_p = 1 - \prod_{i \in p} \exp\left(-\frac{\min(|x_i - a_i^{(1)}|, |x_i - a_i^{(2)}|)}{g_i s}\right), \tag{5.3}$$

where x_i denotes the genomic coordinate in species $i \in p$. Other than in the reference species where this value is given as an input, x_i is determined by linear interpolation between the anchors.

With $d_i = \min(|x_i - a_i^{(1)}|, |x_i - a_i^{(2)}|)$:

$$l_p = 1 - \exp\left(-\frac{1}{s} \sum_{i \in p} \frac{d_i}{g_i}\right). \tag{5.4}$$

The shortest path $\widehat{p}$ through the species graph is then found by minimizing l_p:

$$\widehat{p} = \arg\min_{p \in P} l_p \tag{5.5}$$

with P denoting the set of all paths through G.

IPP's basic function is described as pseudocode in **Algorithm 1** and a simple example schematically visualized in **Figure 5.2**. The algorithm uses max-heap, a specialized tree-based data structure that is an efficient implementation of a priority queue. In a max-heap, the element with the highest priority, in this case the highest projection score, is stored at the root. Two of the basic heap operations are `push`, i.e. adding elements to the heap, and `pop`, i.e. extracting the root from the heap.

5.2.2 SYNTENIC ANCHOR POINT PROPAGATION (SAPP)

Sometimes, instead of projecting single independent genomic locations, we might want to project multiple dependent locations. For example, two enhancers in a syntenic region of the reference species are of course expected to be syntenic in the target species as well. However, **IPP** may render different species paths for both enhancers. Moreover, these two paths might differ in the direction of the closest anchor. Because the **IPP** algorithm estimates the intermediate locations in every bridging species by interpolation, this can lead to projections where the relative positions of the two enhancers are inverted in the target species, neglecting the aspect of synteny (**Figure 5.5 A**).

To that end, I developed a second algorithm termed **SAPP** in which not the genomic location itself but rather its closest anchors are propagated through the species graph. **SAPP** differs from **IPP** in three major aspects. First, when moving from one species to another, the genomic location of interest is not interpolated. Instead, the algorithm follows the anchor points of the initial genomic location. Second, there are no distances between species. Instead, it is only the final anchor span in the target species that is minimized. For example, a potential insertion between the anchors in a bridging species may result in a wide intermediate anchor span, however, this is irrelevant as long as the anchor span in the target species is minimal. Third, anchors are propagated through the species graph independently for both directions. This may implicate resulting target anchors from different species. The **SAPP** algorithm always moves in an outward direction. This characteristic feature of the algorithm ensures that synteny is maintained. Moreover, **SAPP** does not provide point projections.

Algorithm 1 Independent Point Projection

```
function distanceScoringFunction(anchors, coordinate)
    As described in Equation 5.1.
end function

function project(currentSpecies, nextSpecies, currentCoord, currentScore)
    Determine anchors between currentSpecies and nextSpecies that flank currentCoord.
    Estimate nextCoord by linear interpolation between anchors.
    nextScore = currentScore × distanceScoringFunction(anchors, currentCoord)
    return nextScore, nextCoord
end function

function getShortestPath(reference, target, coordinate, speciesList)
    Initialize path as a dictionary with key : value pairs of the form
    species : (score, previousSpecies, coordinate)
    Initialize maxHeap as a heap of tuples of the form (score, species, coordinate).
    maxHeap.push((1.0, reference, coordinate))
    while maxHeap not empty do
        currentBestScore = path[currentSpecies]["score"] if entry exists else 0
        currentScore, currentSpecies, currentCoord = maxHeap.pop()
        if currentScore ⩽ currentBestScore then
            continue                ▷ currentSpecies was already reached by a shorter path.
        end if
        if currentSpecies == target then
            break                                              ▷ Reached target species.
        end if
        for nextSpecies in speciesList do
            nextBestScore = path[nextSpecies]["score"] if entry exists else 0
            if currentScore ⩽ nextBestScore then
                continue            ▷ nextSpecies was already reached by a shorter path.
            else
                args = (currentSpecies, nextSpecies, currentCoord, currentScore)
                nextScore, nextCoord = project(args)
            end if
            if nextScore ⩽ nextBestScore then
                continue            ▷ nextSpecies was already reached by a shorter path.
            else
                path[nextSpecies] = (nextScore, currentSpecies, nextCoord)
                maxHeap.push((nextScore, nextSpecies, nextCoord))
            end if
        end for
    end while
    shortestPathToTarget = Backpropagate path from target to reference.
    return shortestPathToTarget
end function
```

Instead, if the original assumption of maintained synteny is accurate, the resulting anchors in the target species will ultimately span a region that contains all potential orthologs to the corresponding region in the reference species. **SAPP**'s course of action is visually outlined in **Figure 5.5 B** and described as pseudocode in **Algorithm 2**.

Algorithm 2 Syntenic Anchor Point Propagation

```
function moveOut(currentSpecies, currentCoord, currentDirection, target)
    while currentCoord is within outerBoundaries of currentSpecies do
        Walk from currentCoord in currentDirection.
        Stop at the nearest anchor to any other species.
        Set the anchor coordinate in currentSpecies as currentCoord.
        Set the species to which the anchor points as nextSpecies.
        Set the anchor coordinate as nextCoord.
        if nextCoord is within the outer boundaries of nextSpecies then
            if nextSpecies == target then
                Replace respective value in target's outerBoundaries by currentCoord.
            else
                moveOut(nextSpecies, nextCoord, currentDirection, target)
            end if
        else
            Discard anchor and change currentCoord by 1 bp in currentDirection
        end if
    end while
end function

function propagateAnchors(reference, target, coordinate, speciesList)
    for currentSpecies in speciesList do
        Determine the anchors from reference to currentSpecies that flank coordinate.
        In currentSpecies, these anchors span a region (anchorSpan).
        Determine the anchors from currentSpecies to target that flank anchorSpan.
        Set these two anchors as outerBoundaries of currentSpecies.
    end for
    for currentDirection in ["upstream", "downstream"] do
        moveOut(reference, coordinate, currentDirection, target)
    end for
    return outerBoundaries
end function
```

5.2.3 DATA SOURCES AND PROCESSING

DATA SOURCES Zebrafish functional genomics data was downloaded as bam- and bigwig-files from the DANIO-code DCC at https://danio-code.zfin.org/ [242, 243] (whole-embryo 24 hpf **ATAC-seq** [244] and **H3K27me3** [245]). Mouse **H3K27me3** data was downloaded from **ENCODE** for multiple tissues at embryonic stage E10.5 (forebrain, midbrain, hindbrain, heart, limb, embryonic facial prominence) [210], and merged to simulate whole-embryo data comparable to that of zebrafish. Functional genomics data used for enhancer prediction in mouse embryonic limb E10.5 was downloaded from **GEO** (**ATAC-seq** [246], **H3K4me1** and **H3K4me3** [247]) and from **ENCODE** (**H3K27ac** [248]). Functional genomics data for chicken embryonic limb **Hamburger-Hamilton stage 25 (HH25)** [249, 250] was kindly provided by the research group of Stefan Mundlos and a manuscript by Ringel et al. is in preparation, which also provided the locations of putative enhancers in the *Fat1* locus. Putative enhancers in the *Sox9* locus were obtained from Despang et al. [251].

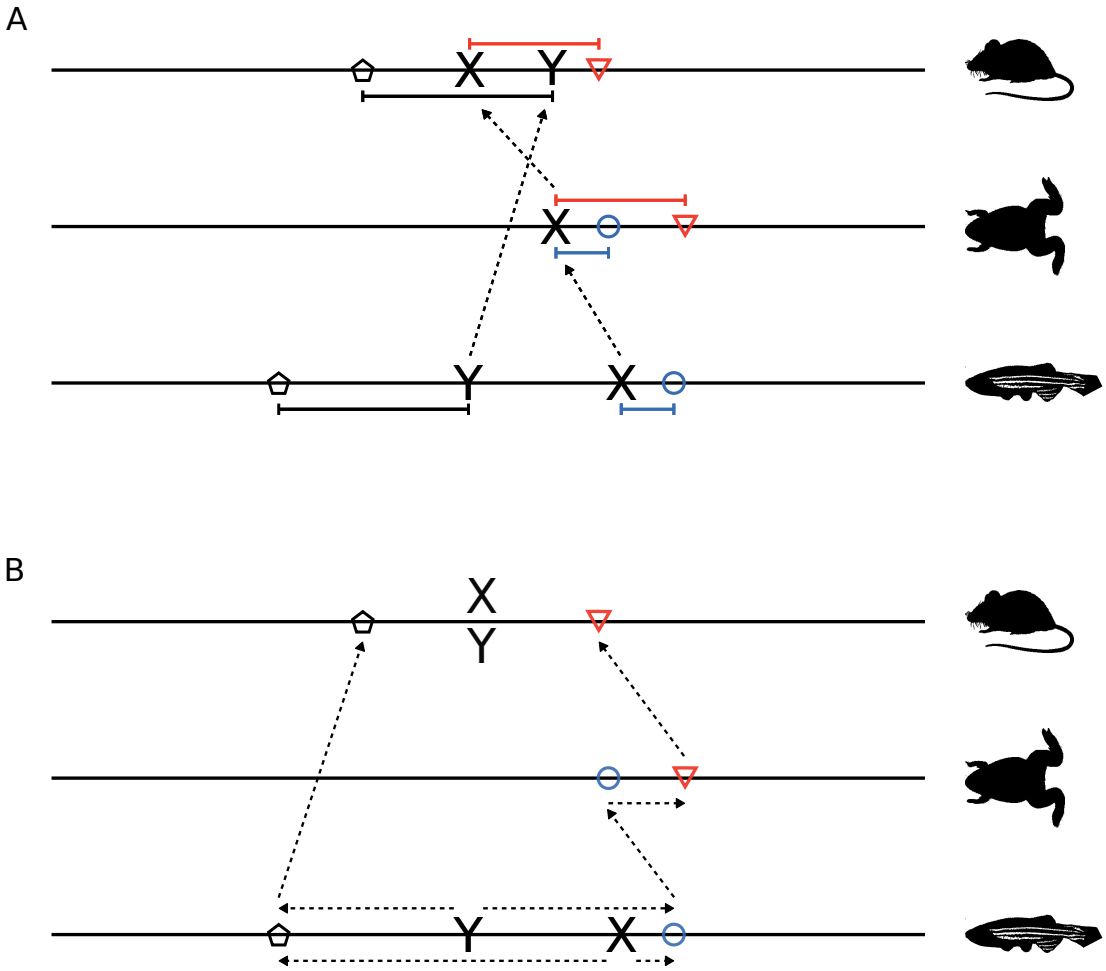

Figure 5.5: Schematic illustration of the **SAPP** algorithm for anchor point propagation from zebrafish to mouse with frog as a bridging species. Symbols represent anchors between two species with their colors and shapes highlighting species pairs (e.g. black pentagons correspond to anchors between zebrafish and mouse). The algorithm starts at a region (X, Y) in the reference species and moves outward in both directions to an anchor to any species until the target species is reached, and choses the path that minimizes the target anchor span. The algorithm is outlined in **Algorithm 2**.

CNES AND GRBS I identified **CNEs** and **GRBs** using CNEr [252]. First, I called direct **GRBs** from zebrafish to mouse. These definitions are vague due to the low **CNE** density between those species. To that end, I called clade-specific **GRBs** (zebrafish - grass carp and mouse - human), and refined the boundaries of zebrafish - mouse **GRBs** by requiring overlap with the clade-specific **GRBs** in both zebrafish and mouse, yielding 152 **GRBs**.

SPECIES SELECTION **IPP** and **SAPP** rely on the assessment of multiple species for close anchor points. A suitable set of bridging species always depends on the particular species comparison. In **Subsection 5.3.4** I map unalignable regions between

zebrafish and mouse. For that, I include bridging species that diverged from mouse later than zebrafish did (human, chicken, frog, coelacanth) and others that diverged from zebrafish later than mouse did (grass carp, common carp, goldfish, medaka, stickleback, japanese pufferfish, green spotted puffer, spotted gar). This increases the probability that a putative functionally conserved element is alignable or close to alignable anchors in some species pairs even though in mouse and zebrafish it is not, and ultimately that those species pairs allow forming a short path through the species graph resulting in accurate projections. In addition, I chose elephant shark as an outgroup for two reasons. First, elephant shark has the slowest evolving genome among all known vertebrates [235]. It therefore resembles the last common ancestor with zebrafish and mouse stronger than any other species in this clade. Second, anchor points shared between elephant shark and zebrafish are likely to be different to those shared between elephant shark and mouse, hence increasing total anchor point density. This follows from the fact that mouse and zebrafish evolved independently and thus accumulated different sets of random mutations. Other slowly evolving species included in the graph are coelacanth and spotted gar, with the latter being the closest relative to zebrafish that did not undergo the additional **whole genome duplication (WGD)** of teleost fish [253]. In **Subsection 5.3.1** I investigate functional conservation of putative enhancers in the developing limb of mouse and chicken. I therefore replaced several teleost species with closer relatives to chicken (penguin, ostrich, alligator, lizard). The respective selections of bridging species are shown as phylogenetic trees in **Figure 5.6**.

FILTER ANCHOR POINTS A fundamental requirement for both **IPP** and **SAPP** is the availability of anchor points with conserved synteny, i.e. neighboring anchor points must be collinear between two species and reside in roughly the same genomic neighborhood. Synteny is an inherent property of **GRBs**, and their genes (both target and bystander) as well as their **CNEs** are largely collinear [110, 233], providing ideal terrain for those methods. In addition to genes and **CNEs**, I decided to use any pairwise alignment to obtain the largest possible sets of anchor points. However, pairwise alignments frequently contain false positives [166], resulting in syntenic regions that include outliers pointing to different chromosomes or simply violating the collinearity constraint. I therefore implemented an outlier-check prior to defining the closest anchors of a region of interest that collects the ten closest anchor points in both up- and downstream direction and discards every data point that fails this test.

5.2.4 IMPLEMENTATION AND AVAILABILITY

I implemented **IPP** and **SAPP** in python and released them together in a repository available at https://github.com/tobiaszehnder/genomic_coordinate_projection.

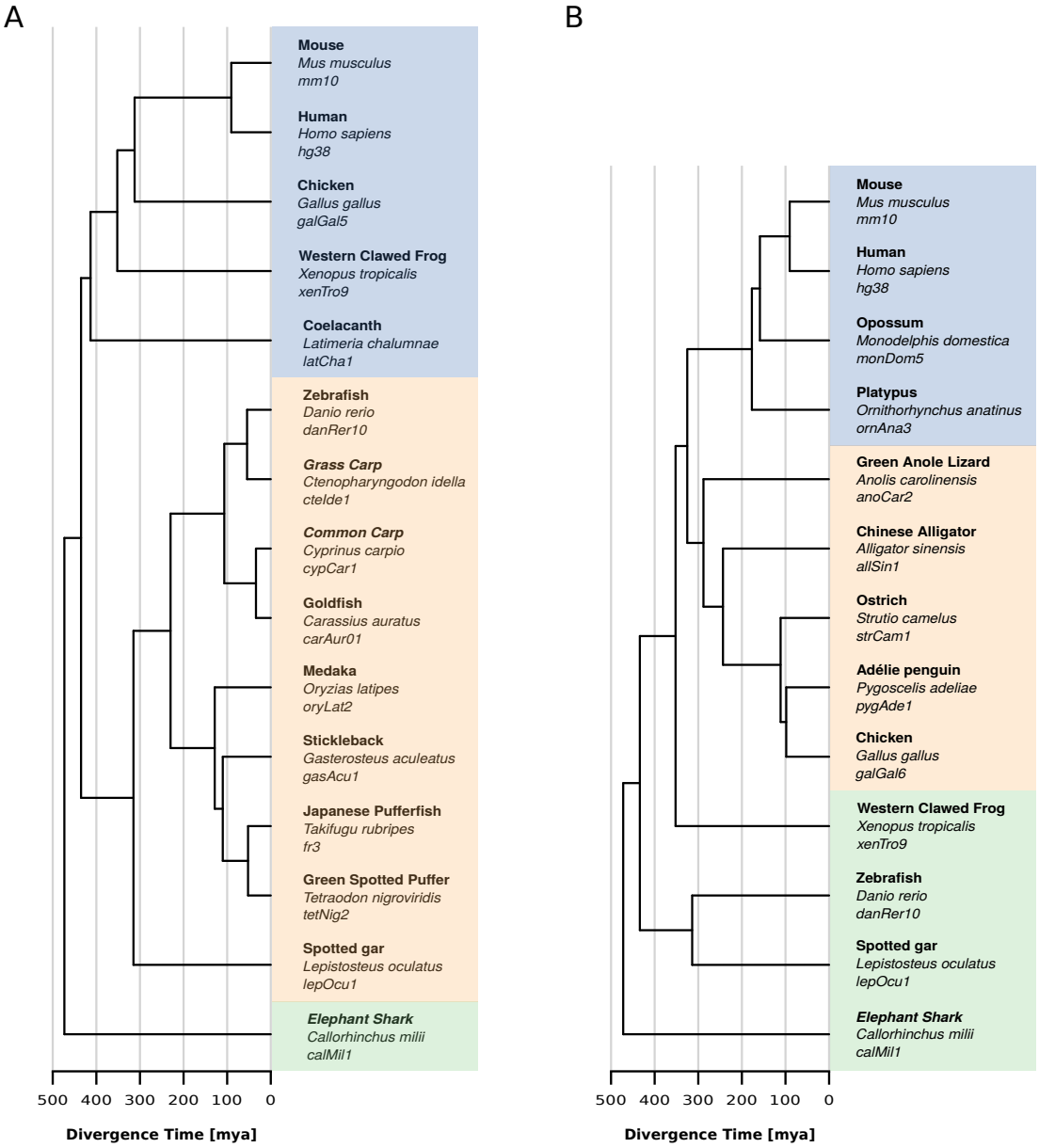

Figure 5.6: Phylogenetic trees of the selected species for the genomic coordinate mapping between mouse and zebrafish (A) and between mouse and chicken (B). Evolutionary divergence times according to Kumar et al. [234]. The labels depict each species' trivial name (bold), scientific name (italic) and genome assembly. Colors indicate monophyletic groups. **A** blue - Sarcopterygii, (including tetrapods and lobe-finned fish such as coelacanths and lungfish), orange - Actinopterygii (including teleost, gars and bowfins), green - outgroup. **B** blue - Mammalia, orange - Sauria (including reptiles and birds), green - outgroup.

5.3 RESULTS

5.3.1 IDENTIFYING FUNCTIONAL ORTHOLOGS USING IPP

Key developmental genes often reside in tightly regulated neighborhoods with relatively high frequencies of sequence conservation. At the same time, many putative enhancers in these genomic environments lack direct alignability between a given species pair, especially if these species are not closely related. However, as introduced earlier in this Chapter, some of the unalignable enhancers might still share function even though their respective sequences have changed to a degree that has rendered them unalignable. Here I will describe an attempt at identifying putative functional orthologs using the previously described **IPP** method. For that, I used the enhancer prediction method **eHMM** I presented in **Chapter 4** for identifying putative enhancers in the developing limb in mouse, resulting in a total of 4569 predicted enhancers genome-wide. I used **IPP** to project the genomic locations of these putative enhancers to chicken. Using **IPP**'s projection score (**Subsection 5.2.1**), I grouped the enhancers in three different classes: *S* (sequence-conserved), *F* (potentially functionally conserved) and *N* (not conserved). Enhancers were classified as *S* if their projection score using only direct alignments from mouse to chicken was above a certain threshold, and as *F* if they were not in *S* and their projection through the species graph yielded a score greater than the threshold in question. I set the projection score thresholds to 0.99 and chose a stringent distance half life of 10 **kbp** for determining the scaling factor (**Equation 5.2**). According to that and **Equation 5.1**, a projection score of 0.99 corresponds to an enhancer-anchor distance of ~150 **bps** when considering direct anchors from mouse to chicken. The projections were performed on the enhancers' center **base pair** coordinate, and given a typical enhancer's width of a few hundred **bps** [171], a distance of less or equal than 150 **bps** to an anchor can be considered as at least partially overlapping the enhancer. However, enhancers in class *F* were projected using at least one bridging species, and since the distances in the exponent of the score are additive, this means that the *sum* of all distances to the anchor points of each used species is less or equal to the corresponding distance, e.g. 150 **bps**.

According to these classifications, 393 (9%) of the predicted enhancers have a conserved sequence between mouse and chicken, and another 567 (12%) are identified as candidates for functional conservation as they overlap indirect alignments via at least one bridging species (**Figure 5.7 A**). *N* enhancers are generally lowly conserved in a multiple species comparison as depicted by the distribution of phastCons 60way scores in **Figure 5.7 B**. This score is computed from a multiple alignment across 60 vertebrates and correlates with the fraction of species in which an enhancer is conserved. Conversely, enhancers in class *S* exhibit more variable conservation scores

with a larger fraction not only conserved to chicken but also to many other vertebrates. The distribution of the conservation scores of *F* largely resembles that of *S*, reflecting the identified alignments between the species pairs in the respective optimal bridging species paths. In terms of conservation across multiple vertebrates, *F* enhancers are hardly distinguishable from *S* enhancers despite their missing direct alignment between mouse and chicken.

I analyzed the projected genomic locations for the presence of epigenetic features associated with enhancer function such as chromatin accessibility measured by **ATAC-seq**, **histone modifications H3K27ac**, **H3K27me3**, **H3K4me1** and **H3K4me2**, as well as binding events of **RNAP II** (**Figure 5.7 C**). Sequence-conserved enhancers from class *S* possess the strongest enhancer-associated features whereas the projections of non-conserved enhancers point to regions in chicken that are much less enriched for those. *F* enhancers generally exhibit these features similar to *S* and rather distinct from *N*, suggesting that these regions might indeed act as enhancers in chicken, too.

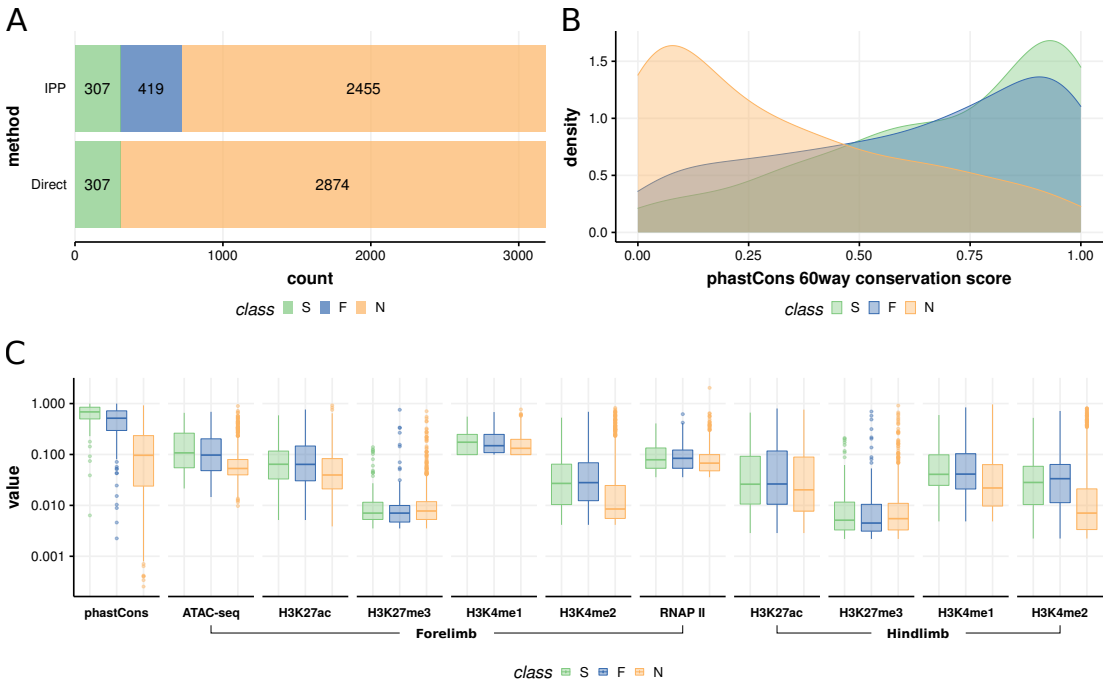

Figure 5.7: Identification and evaluation of potential functional orthologs between mouse and chicken using putative enhancers predicted by **eHMM**. **A** Enhancer classification into sequence-conserved (*S*), potentially functionally conserved (*F*) and not conserved (*N*) enhancers according to a projection score threshold of 0.99. Elements in F have a score above the threshold for projections using multiple bridging species, elements in S have a score above the threshold for projections using only direct alignments between mouse and chicken. **B** Mouse mm10 phastCons 60way sequence conservation scores averaged over 500 **bps** windows centered on the enhancers. **C** Distribution of the signal of phastCons and various functional genomics experiments averaged over 500 **bps** windows centered on the IPP projections in chicken **HH25**.

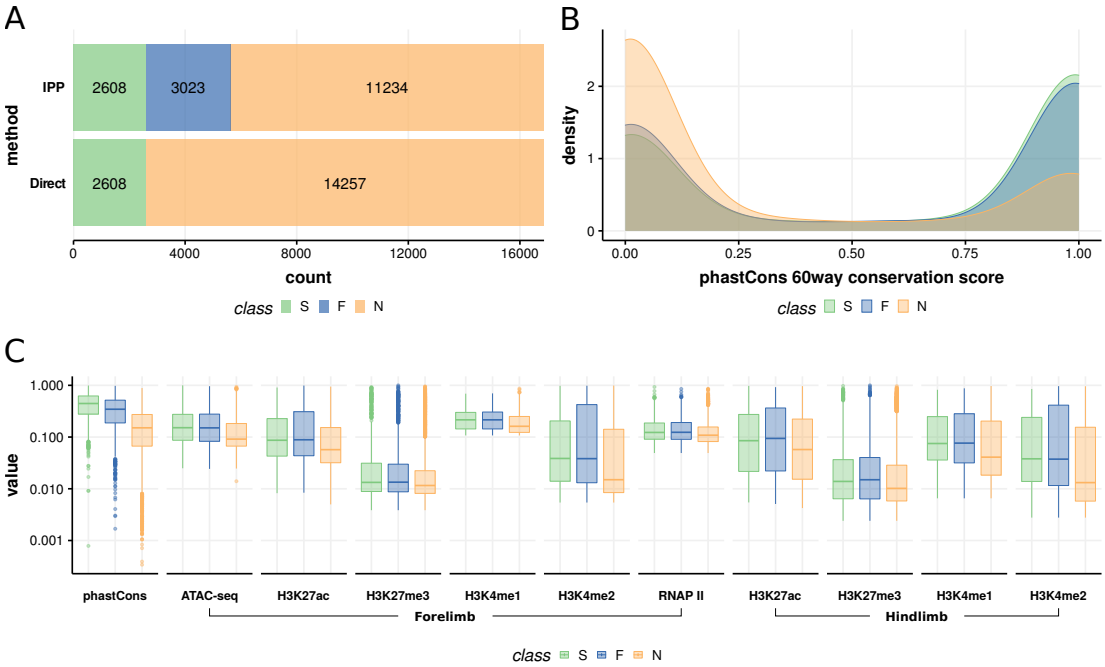

Figure 5.8: Identification and evaluation of potential functional orthologs between mouse and chicken using **ATAC-seq** peaks predicted by MACS2 [207]. Subfigures are analogous to **Figure 5.7**.

As described in **Chapter 4**, **eHMM** predicts enhancers with a particular emphasis on specificity rather than sensitivity. The number of enhancers assessed for potential functional conservation is therefore rather low. For this reason I expanded the query to a larger set of putative regulatory elements, namely all **ATAC-seq** peaks predicted by MACS2 [207] in mouse forelimb that are located within a **GRB**, resulting in a total of 16,865 elements. The distribution of element classes is shifted towards conservation, with *S* and *F* amounting to 15% and 18% of all tested elements, respectively (**Figure 5.8 A**). This is roughly twice as many conserved elements as observed for the predicted enhancers and arises from the fact that **ATAC-seq** peaks include active promoters which are often conserved. In terms of sequence conservation across 60 vertebrates and the functional genomics features in chicken, the **ATAC-seq** peaks show similar results to the predicted enhancers with respect to the classes *S*, *F* and *N*. Overall, **IPP** identifies more than 3000 candidates for potential functional conservation with substantiating epigenetic features in the developing limb bud.

These results demonstrate global conservation patterns and suggest the prevalence of functional conservation throughout the genome. Single loci, however, might differ from the global average. To that end, I examined conservation of regulatory elements in five limb-specific loci, i.e. the **GRBs** surrounding the limb-specific genes *Fat1*, *Fgf8*, *En1*, *Hand2* and *Sox9*. Conservation levels differ remarkably between different loci (**Figure 5.9**). For example, half of the **ATAC-seq** peaks in the **GRB** around the gene

Sox9 are either sequence-conserved or potential candidates for functional conservation, whereas in the *Fat1* **GRB**, 90% of the **ATAC-seq** peaks are not conserved. An overview of the features for the individual loci is given in **Figure 5.10**.

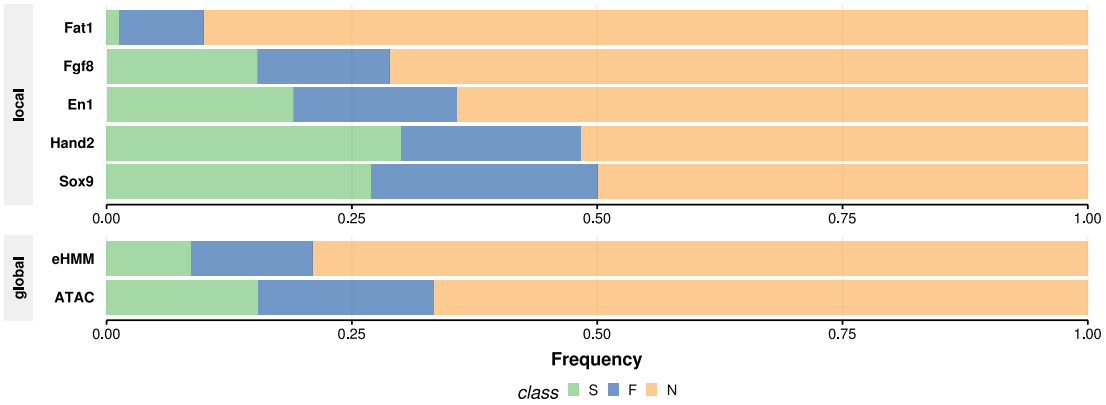

Figure 5.9: Local and global enhancer classification into sequence-conserved (*S*), potentially functionally conserved (*F*) and not conserved (*N*) enhancers according to a projection score threshold of 0.99. Label names of local classifications refer to the **GRBs** in which the respectiv genes reside. The labels **eHMM** and **ATAC** refer to global classifications using elements from the whole genome as predicted by **eHMM** and MACS2, respectively.

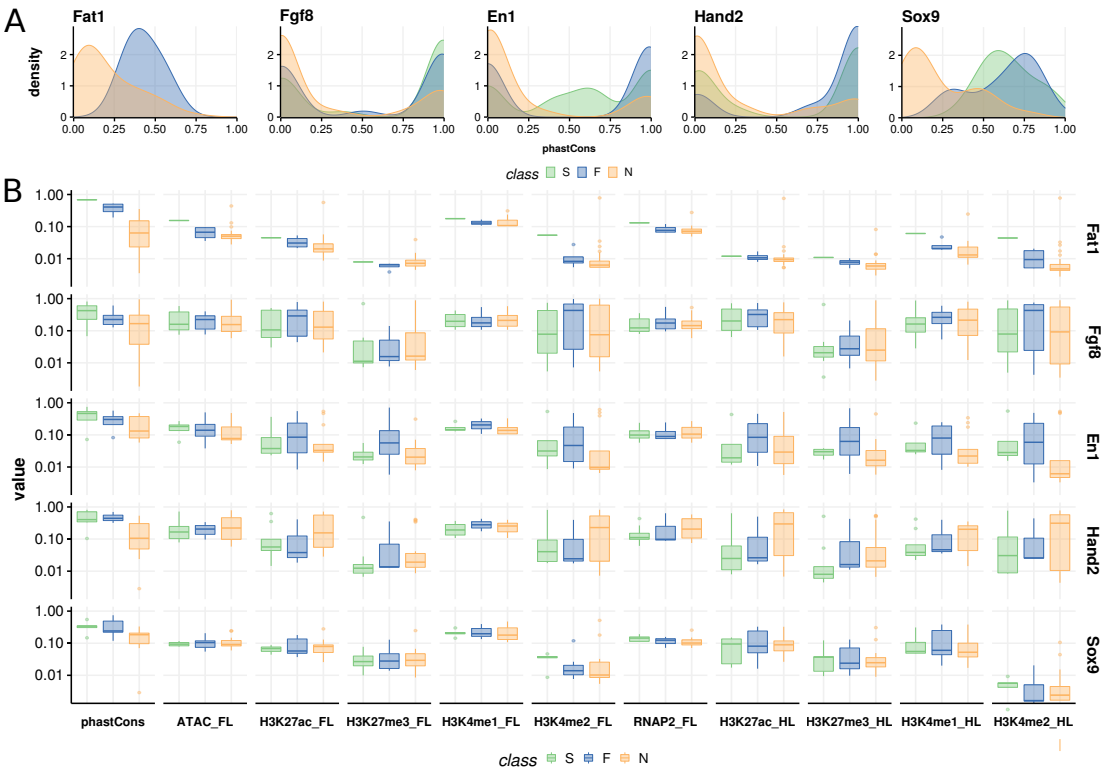

Figure 5.10: Potential functional orthologs between mouse and chicken in the **GRBs** encompassing *Fat1*, *Fgf8*, *En1*, *Hand2* and *Sox9*. **A** PhastCons sequence conservation scores averaged over 500 **bps** windows centered on the **ATAC-seq** peaks. **B** Feature distributions of the **IPP** projections in chicken.

OPTIMAL PATH STATISTICS **IPP** optimizes the choice of bridging species for maximizing projection accuracy. **Figure 5.11** shows the distribution of path lengths, i.e. the number of bridging species for the classes *S*, *F* and *N* and shows the relative frequencies of the different species occurring in those paths for a conservation threshold of 0.99. Optimal paths pass through up to six bridging species with shorter paths being more common. Especially shorter paths are predominantly going through mammalian species human, opossum and platypus with a strong favor for human in paths with a single bridging species. Compared to the mammalian clade, species from the clade Sauria (including reptiles and birds [254]) are less frequently used as a bridging species in shorter paths and approximately as often in longer paths (e.g. $\geqslant 3$). Species from the outgroup are the least frequent and mainly appear in approximately 5% of the shorter paths.

Elements classified as sequence-conserved (*S*), i.e. a direct alignment from mouse to chicken is located within roughly 150 **bps**, are projected directly in only about 50% of the cases. The other half of elements in *S* are considered directly sequence-conserved to mouse, too, yet they were projected through paths comprising up to 6 species similarly to elements from the other classes. Within a clade, the minimal evolutionary distance between a species to either mouse or chicken is the strongest determinant of its employment as a bridging species. Hence, human is more prevalent than opossum which in turn surmounts platypus among the mammals, and xenopus is the most frequently used outgroup species. The exception is Sauria, where alligator altogether appears more often as a bridging species relative to the two birds ostrich and penguin, despite its evolutionary trajectory having diverged from birds more than 100 **million years (myr)** earlier than ostrich and penguin diverged from chicken.

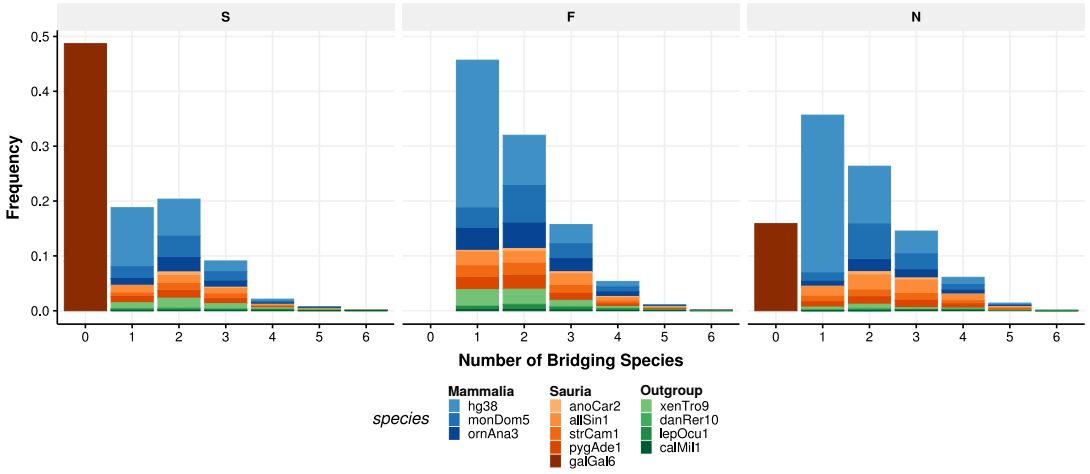

Figure 5.11: Bridging species frequencies in projection paths for the classes *S*, *F* and *N* by path length.

5.3.2 QUALITATIVE EVALUATION OF IPP'S PROJECTION QUALITY

In this Subsection I want to demonstrate the benefit of **IPP**'s projections over the *Direct* approach in a qualitative assessment of an example locus roughly 100 **kbp** around the gene body of *Wbp1l* in the **GRB** of *Fgf8* in mouse. **Figure 5.12** depicts the respective genomic regions in mouse and chicken. Matching alignments are connected with lines and show two alignment-dense regions as well as a central 'alignment-desert' with no direct alignments for a stretch of approximately 55 **kbp**. I projected three putative regulatory elements (**A1**, **A2**, **A3**) from mouse to chicken using **IPP** and the *Direct* approach. **A1** is located right next to a direct alignment between mouse and chicken. The projections of **IPP** and *Direct* therefore agree and have both close to maximal scores. **A2** and **A3**, however, reside in the alignment-desert. Consequently, the *Direct* projections are uncertain with scores of 0.54 and 0.24 for **A2** and **A3**, respectively. **IPP**'s projections benefit from a higher anchor point density to bridging species and consequently exhibit higher projection scores of 0.68 and 0.97. While the accuracy of a projection with a score of 0.68 (**A2**) is still rather uncertain, the projected location of **A3** with a score of 0.97 promises to be highly accurate. Indeed, the projection of **A3** overlaps with epigenomic features that are associated with enhancers, i.e. enriched **ATAC-seq**, **H3K27ac**, **H3K4me1** and **H3K4me2** and agrees with the features in mouse, suggesting that **A3** is functionally conserved in absence of sequence conservation. **IPP** is able to discover such occurrences even when the closest direct anchors are far apart, in this case more than 50 **kbp**.

5.3.3 QUANTITATIVE EVALUATION OF IPP'S PROJECTION QUALITY

In **Subsection 5.3.1** I have presented an application of **IPP**, a method for the projection of non-alignable genomic elements, in order to identify putative functional orthologs beyond sequence conservation between mouse and chicken. In the next Subsection I will address the question of functional conservation not only on the level of single independent elements, but for dependent neighborhoods with conserved synteny. This work originated from a collaborative project that studies zebrafish as an alternative model organism to mammals and thus researches the comparability of genomics in zebrafish and mouse. In order to assess **IPP**'s suitability for the zebrafish - mouse comparison, I comprehensively evaluated the projection quality of **IPP** by extending the scope from putative functional elements to entire **GRBs**. I projected 152 **GRBs** divided into 1 **kbp** windows from mouse onto zebrafish genomic coordinates. This resulted in a total number of 70,949 bins, thus providing a good opportunity to evaluate **IPP**'s overall projection quality.

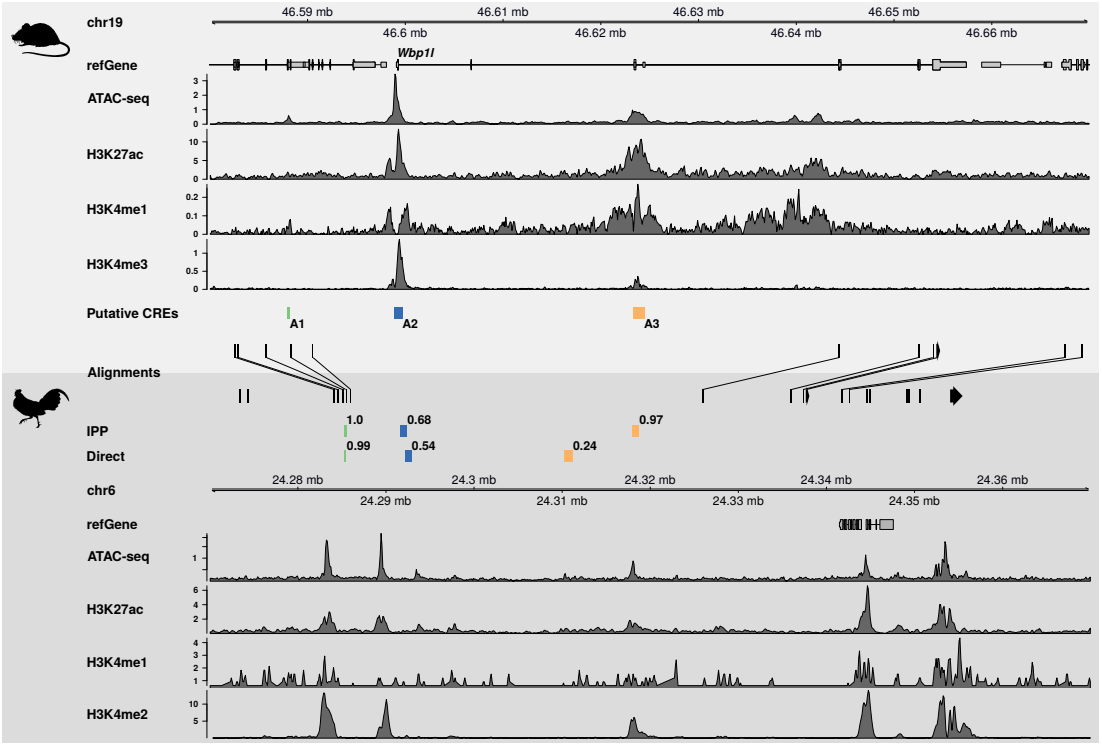

Figure 5.12: Projections of three putative regulatory elements in and around the gene body of *Wbp1l* from mouse (top, light gray) to chicken (bottom, dark gray) using **IPP** and *Direct*. Projection scores are indicated next to the projected elements. The difference in the sets of features between mouse and chicken is according to data availability (**H3K4me3** and **H3K4me2**).

Using **IPP** instead of interpolating the signal using direct alignments (henceforth referred to as the *Direct* method) highly increased the projection accuracy. The projection accuracy is reflected by the distance of a particular genomic coordinate to its closest anchor point. The closer an anchor point, the more confident the projection. **Figure 5.13 A** depicts the distributions of distances from the original coordinates to their closest anchors in zebrafish (left panel) and from the projected coordinates to their closest anchors in mouse (right panel). In zebrafish, **IPP** decreases the median distance from 4.8 **kbp** using direct alignments to 0.5 **kbp**. In mouse, the distances average to 8.7 **kbp** (*Direct*) and 0.6 **kbp** (**IPP**). The fraction of coordinates located within only a few **bps** to an anchor point is remarkably increased in both zebrafish and mouse, indicating that many coordinates were indirectly alignable through a particular combination of bridging species using **IPP**. In zebrafish, 12,636 regions lie within 10 **bps** to their closest **IPP** anchor compared to 1513 using *Direct*. In mouse, this effect is even stronger with 20,017 (**IPP**) vs. 1892 (*Direct*) regions within 10 **bps** to their closest anchor.

Assessing distances to the closest anchors in the reference and target species is often a good approximation of the projection accuracy, however, it can be deceptive. For

example, the projection of a region close to its anchor points in both zebrafish and mouse can still be uncertain if the projection passed a bridging species where its intermediate projection and the respective anchor were more distant. I therefore additionally evaluated projection accuracy by assessing the projection score (**Subsection 5.2.1**) which reflects the distances from all intermediate coordinates to their respective anchors instead of only one species at a time. It is therefore a more stringent measure. **IPP** projections are distinctly shifted towards high projection scores (**Figure 5.13 B**, mean projection scores 0.82 and 0.61 for **IPP** and *Direct*, respectively). As a result, the **GRBs**' anchor point density in zebrafish was increased approximately 5 times using **IPP** (77,542 unique anchors) compared to the *Direct* approach (15,297 unique anchors). This is reflected by the distribution of inter-anchor distances in **Figure 5.13 C** and in an example locus in **Figure 5.14**, colored bars. On average, **IPP** anchors are separated by 1.0 **kbp**, *Direct* anchors by 5.2 **kbp**. Particularly the tail towards large distances representing anchor deserts is lost when optimizing the selection of bridging species. Instead, the second mode of distances within direct anchors representing inter-cluster distances as discussed in **Section 5.1** is reduced, suggesting that anchor points are more prevalent between those clusters. Moreover, **IPP** anchors are enriched for very short inter-anchor distances, indicating that also intra-cluster anchor densities are increased. Together, these results affirm an increase in projection accuracy of **IPP** over the *Direct* approach and endorse its application for zebrafish - mouse comparisons.

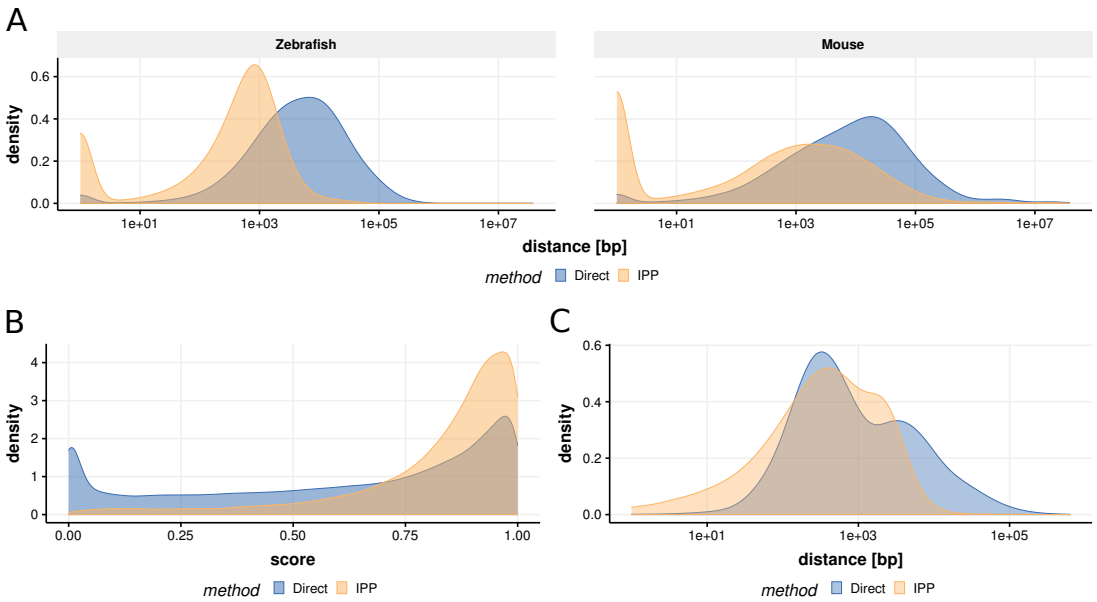

Figure 5.13: Measures of projection quality of **IPP** versus *Direct*. **A** Distributions of distances from the original genomic coordinate in zebrafish to the closest anchor (left panel), and from the projection to the closest anchor in mouse (right panel). **B** Distribution of projection scores. **C** Distribution of distances between anchors in zebrafish reflecting anchor point density.

5.3.4 CONSERVATION OF TOPOLOGICAL CHROMATIN STRUCTURES IN ABSENCE OF SEQUENCE CONSERVATION

The remarkable syntenic arrangement of **CNEs** within **GRBs** has garnered attention in the scientific community towards the evolutionary constraints that led to this observation. A well-established hypothesis assumes that keeping the regulatory elements in cis contributes towards conservation of synteny [255]. Moreover, the distribution of the regulatory elements i.a. in gene bodies of bystander genes adds additional constraints to the location of **CNEs**. Recent reports about the co-localization of **GRBs** and **TADs** have added another perception and suggested a regulatory role of **CNEs** towards the three-dimensional structure of chromatin [112]. A possible mechanism by which **CNEs** can influence chromatin structure is through the interaction with **Polycomb (PcG)** and **Trithorax (TrxG)** group proteins, two antagonistic groups of chromatin-modifying factors. **PcG** proteins are associated with transcriptional repression [256, 257], **TrxG** proteins with its promotion [258]. **Polycomb repressive complex (PRC)** 1 and 2 are two major multiprotein complexes from the family of **PcG** proteins. **PRC2** deposits the **histone modification H3K27me3** [259, 260] which can then be read by **PRC1** which effects chromatin compaction [261]. In addition, **PcG** proteins are associated with the stabilization of chromatin loops and chromatin domains [262]. **H3K27me3** is a relatively stable chromatin mark that is maintained throughout replication and inherited to the daughter cells, where it then needs the activity of **PcG** proteins to proliferate [263]. Because of this epigenetic memory and together with the described effects of **PcG** proteins on chromatin topology, **H3K27me3** might therefore serve as a potential readout for established three-dimensional chromatin structures. Similarities of epigenomic features between species suggest functional conservation of intra-**TAD** subdomains and enhancer topology organisation.

To test the hypothesis of conserved chromatin topology in absence of sequence conservation, I mapped the signal from **H3K27me3 ChIP-seq** experiments from mouse onto zebrafish coordinates using **IPP**. To that end, I limited the regions of interest to **GRBs** and divided them into 1 **kbp** windows which I then mapped between zebrafish and mouse using **IPP**. I assessed the projection quality in the previous **Subsection 5.3.3**. In the remaining Subsection I will assess the extent to which the signals overlap between the two species and address the question of functional equivalence.

TOPOLOGICAL SUBDOMAINS OF CONSERVED EPIGENOMICS With the confidence of **IPP** producing high accuracy projections as demonstrated in **Subsection 5.3.3** I compared the **H3K27me3** signal between mouse and zebrafish using **IPP** to map the binned **GRBs**. For that, I used data from matched developmental stages 24 hpf in ze-

brafish and E10.5 in mouse. I quantile normalized the signal distributions in the two species and investigated signal overlap. Using the 90th percentile of the **H3K27me3** signal as an enrichment threshold I classified the **GRB**-bins into three types of subdomains. Subdomain α contains regions with mutually enriched **H3K27me3** in both zebrafish and mouse. β comprises regions with mutually depleted **H3K27me3** and δ holds the regions that are differentially enriched in either zebrafish or mouse and depleted in the other. **Figure 5.14** shows the **H3K27me3** mapping and the subsequent classification into the subdomains α, β and δ in the *Irx* locus. The example shows a region of conserved **H3K27me3** that spreads beyond the immediate promoter region and gene body of *Irx3a* to a region of approximately 90 **kbp** downstream of *Irx3a* (**Figure 5.14**, shaded area), which predominantly classifies as subdomain α. Note that this region is vastly depleted of directly alignable anchors between zebrafish and mouse, yet the epigenomic signal is conserved.

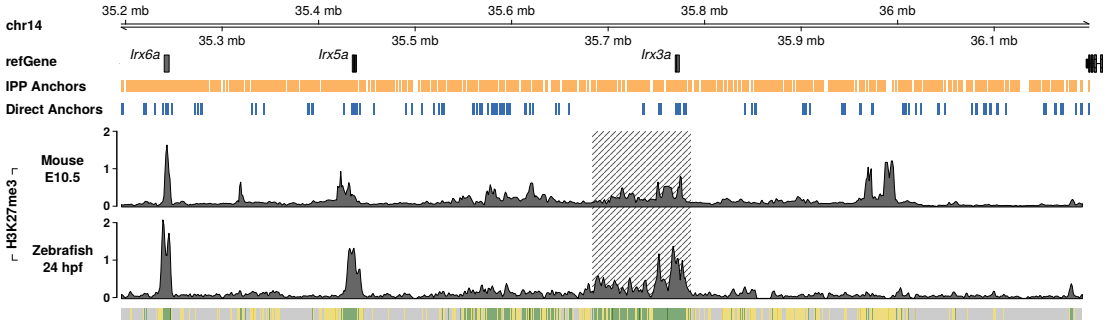

Figure 5.14: Mapping of **H3K27me3** in the *Irx* locus from mouse onto zebrafish genomic coordinates using **IPP** with multiple bridging species. Mapped data was binned into 1 **kbps** windows. Anchor points are indicated as colored dashes. Identified subdomains are indicated as colored bars below the zebrafish track and refer to the following color code regarding the enrichment of the epigenetic signal: green - conserved (α), light gray - depleted (β), yellow - differential (δ).

The global distribution of subdomains is depicted in **Figure 5.15 A**, showing that a total of 2.286 **Mbp** within **GRBs** exhibits epigenomic conservation. α domains are on average wider than δ subdomains while β subdomains are often substantially larger, owed to the fact that low **H3K27me3** is the default state to be expected at random genomic locations (**Figure 5.15 B**). The frequent interspersion of α and δ subdomains further contributes to smaller consecutive domains. The bulk of the α subdomains span less than 10 **kbp** and thus clearly range in a sub-**TAD** scale. Naturally, epigenomic conservation between two species is more likely if the sequence is conserved too. Analogously, many α subdomains contain alignable regions. For example, the conserved genes *Irx3a*, *Irx5a* and *Irx6a* all reside in subdomain α (**Figure 5.14**). However, epigenomic conservation is not restricted to sequence-conserved regions only but also observed when sequences have diverged. This is quantified by the phastCons 12way conservation score of **ATAC-seq** peaks across the three types of

subdomains as shown in **Figure 5.15 D**. α subdomains harbor the biggest fraction of highly conserved elements, however, they do comprise a notable amount of zebrafish-specific elements with very low phastCons scores. 15.2% of all **ATAC-seq** peaks in α have a phastCons score of less than 0.05 (33.5% and 25.6% in β and δ, respectively).

The directionality index calculated from contact frequencies of Hi-C experiments allows identifying boundaries of topological chromatin domains on a sub-**TAD** level [264]. It represents the degree of up- and downstream biases of physical interactions between genomic loci, and this bias is maximal at domain borders where loci interact exclusively with loci from one direction [16]. Hence, the directionality index' absolute value is a readout for a locus' proximity to a domain border independent of direction. **ATAC-seq** in α domains are more likely to be situated at subdomain boundaries as they exhibit a light shift in the distribution of the absolute directionality index towards higher values (**Figure 5.15 C**). Enhancers at domain boundaries have been reported to be involved in the regulation of multiple targets in a process called loop extrusion [265, 266]. However, the effect size is relatively small, suggesting that this property pertains to only a small subset of those elements. In the next paragraph I will investigate whether those elements experiencing a conserved epigenetic influence effectively show signs of functional equivalence that may have had an impact on the epigenomic landscape and thus ultimately on the higher order chromatin topology.

FUNCTIONAL EQUIVALENCE OF ENHANCERS WITHIN TOPOLOGICAL SUBDOMAINS

The task of investigating functional equivalence of enhancers across species requires the correct assignment of potentially equivalent elements. On average, **IPP** projects genomic coordinates with high accuracy. However, it does so for point coordinates, and finding a matching enhancer based on point projections including a potential level of uncertainty is error-prone and not trivial. For example, if **IPP** projects an enhancer from zebrafish to mouse into the vicinity of multiple enhancers, it might not be trivial to choose an appropriate distance limit in order to decide which enhancers to include as potential candidates for equivalence.

Instead of projecting point coordinates, **SAPP** propagates an enhancer's anchor points while respecting the constraint of conserved synteny such that the resulting anchor span in the target species is minimized. This yields a genomic region of minimal width in which a potential equivalent enhancer will be located given that synteny is conserved between the reference and the target species and given that an equivalent enhancer exists. It is therefore the more suitable approach for narrowing the search field for candidates of functional equivalence.

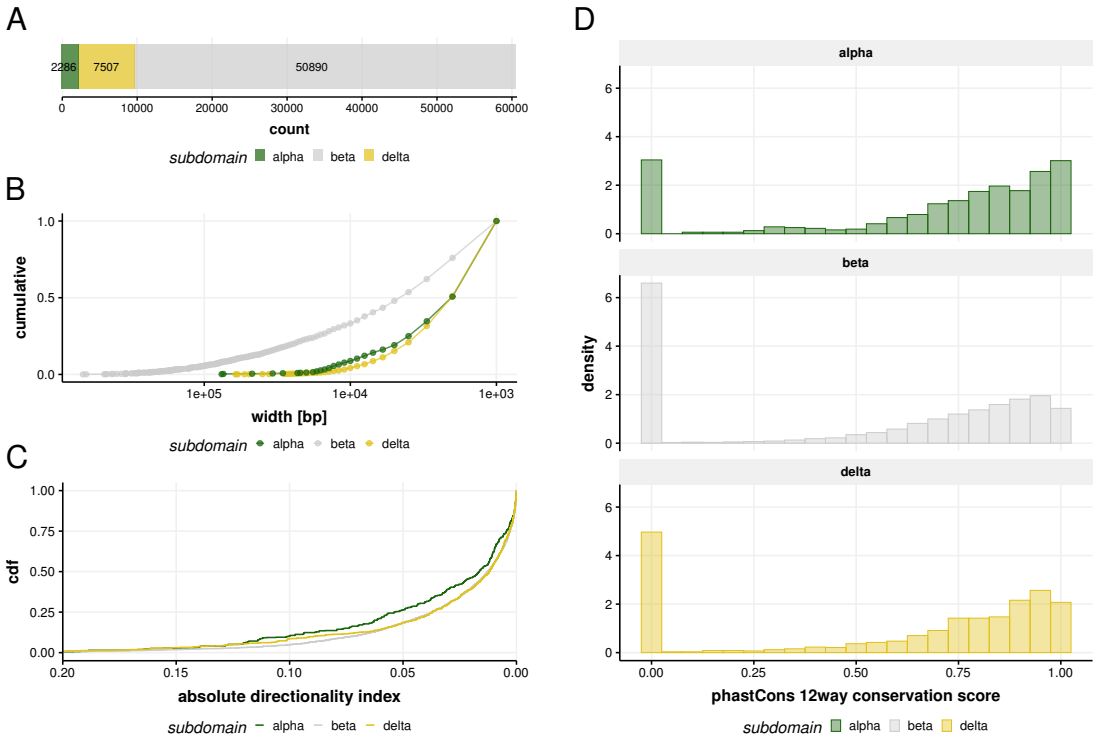

Figure 5.15: Characterization of identified subdomains. **A** Number of bins classified as subdomain α, β and δ. **B** Cumulative distribution of the widths of consecutive subdomains. **C** Cumulative distribution of the absolute value of the directionality index of **ATAC-seq** peaks within the subdomains. **D** Distribution of phastCons 12way conservation scores of **ATAC-seq** peaks within the subdomains.

I propagated the anchors of 9430 **ATAC-seq** peaks within **GRBs** using **SAPP**, resulting in narrower anchor spans in both the reference and the target species compared to the direct anchors from zebrafish to mouse (*Direct*). The distribution of the anchor span widths in both species is depicted in **Figure 5.16**. The median width of the anchor spans using the *Direct* approach is 24.0 **kbp** and 47.5 **kbp** in zebrafish and mouse, respectively. In contrast, the median of the resulting anchor spans using **SAPP** are 7.1 **kbp** (zebrafish) and 14.6 **kbp** (mouse), corresponding to more than a three-fold reduction of the average anchor span width. In addition, **SAPP** substantially increases the fraction of anchor spans with a width in the order of only a few hundred **bps** both in the reference and the target species, corresponding to regions that are located on alignments throughout the path from the reference to the target. Hence, they are indirectly alignable w sith the correct choice of bridging species. The reason that these regions, which are located on an alignment, do not have an anchor span width of zero lies in the nature of **SAPP**'s design, according to which regions overlapping an alignment get assigned the alignment's start and end coordinates as their anchor points. By that, **SAPP** considers the possibility of gaps in the alignments.

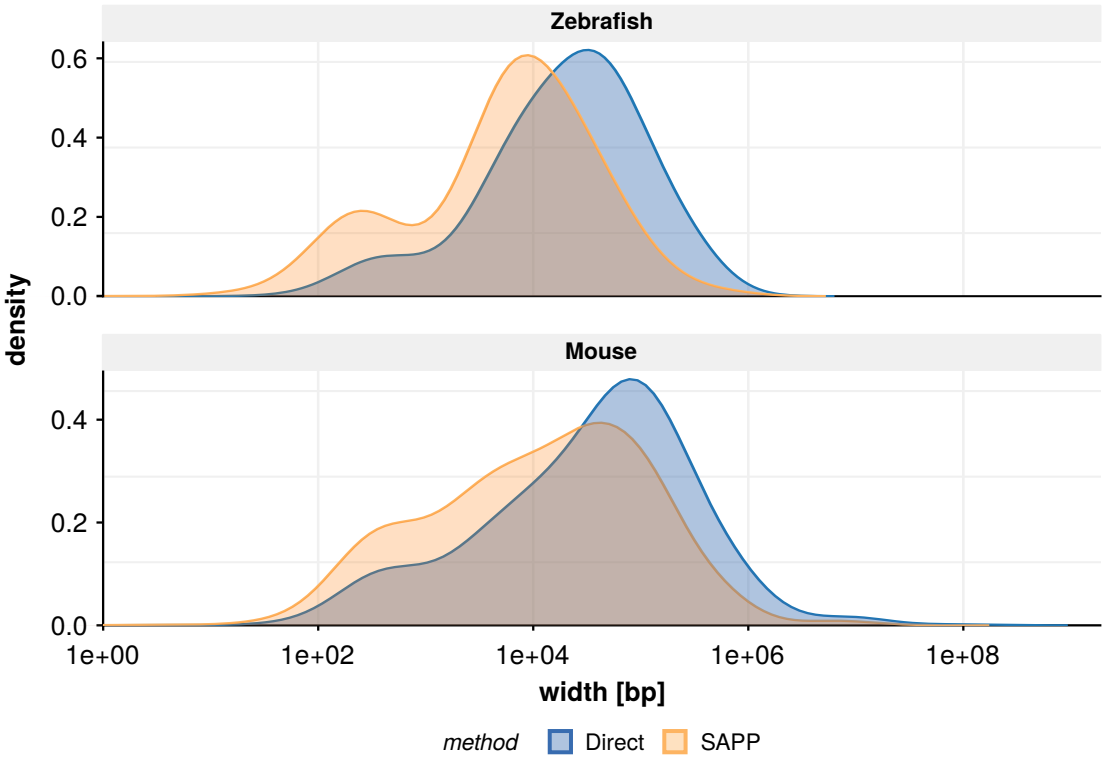

Figure 5.16: Projection quality as measured by the distribution of anchor span widths in zebrafish and mouse using **SAPP** versus *Direct*.

Narrowing the resulting anchor spans in the target species as much as possible is pivotal for the purpose of identifying functional equivalence in syntenic regions. The target anchors of some of the assessed elements are still fairly distant and in those cases an accurate assignment of equivalent elements might still be difficult. To that end, I reduced the set of zebrafish elements to those with target anchor span widths of less than 1 **kbp**, yielding 8360 candidate elements. It is worth noting that the *Direct* approach only yielded 1096 elements with the same threshold. Naturally, a large part of those elements may be truly specific to zebrafish or teleost fish and may indeed lack conservation to mouse as those two species diverged over 435 **mya** As a consequence, only 1117 of the 8360 candidates actually have a complementary candidate element (i.e. a **DNase-seq** peak) within the determined anchors.

To test for functional equivalence I examined the two species' elements within shared subdomains for the number of pairwise matching **TFBS** motif hits according to seqpattern (threshold 0.85) [267], using 190 **TF** motifs from JASPAR [268] and TRANSFAC [269]. As a control, I counted the number of matching motifs to an equally sized control set of randomly picked elements from non-matching domains in cis and trans. Alignable elements are often conserved promoters and it does not surprise that they are more likely to have matching motifs across species independent of the type of their allocated subdomain (**Figure 5.17**). This effect is present in non-alignable ele-

ments, too, albeit to a less pronounced extent in subdomain β and δ. Non-alignable elements in α, however, resemble the alignable elements in that they are more homotypic between zebrafish and mouse than expected. It should be noted that the sample size in some groups is rather small, impeding the generalization of these results.

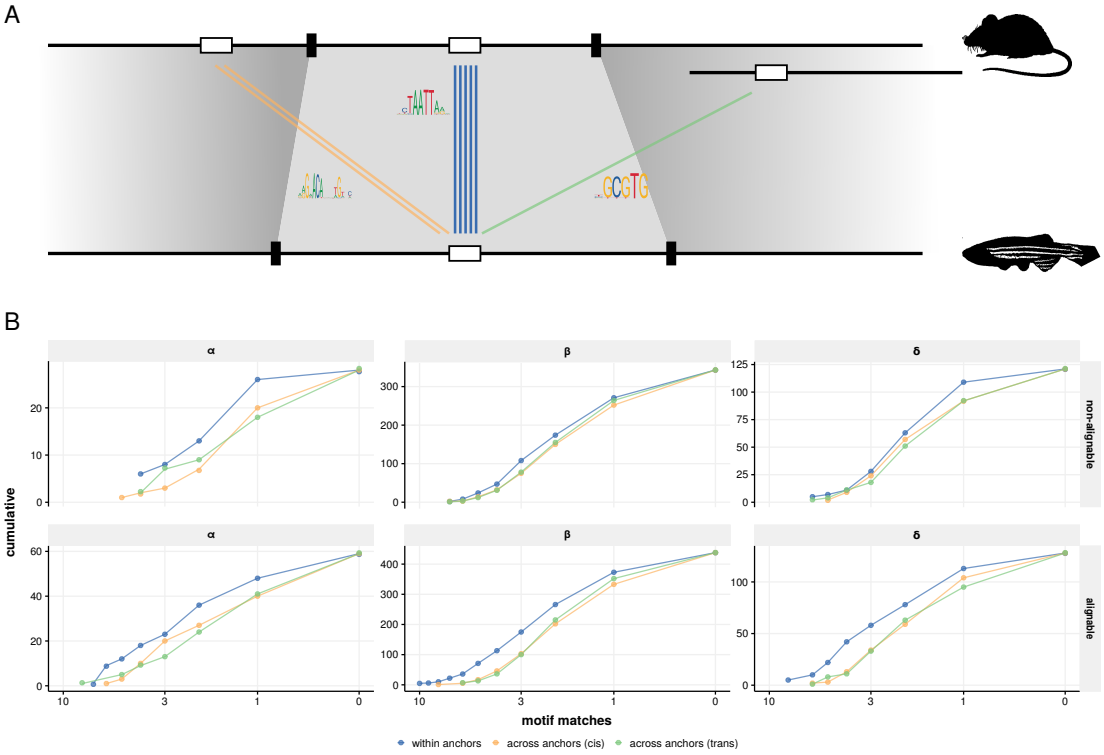

Figure 5.17: TFBS motif match analysis. **A** Schematic illustration of the motif match analysis. Zebrafish and mouse elements within anchor boundaries (blue links) are assessed for the number of shared motifs and compared to randomly picked controls from outside the boundaries in cis (orange) and trans (green). **B** Cumulative distributions of the number of shared motifs between elements within and across anchor boundaries.

5.4 DISCUSSION

In this Chapter I investigated functional conservation of cis-regulatory elements in absence of sequence conservation as an orthogonal approach to the traditional perception of conservation solely based on sequence. In **Subsection 5.2.1** I introduced the notion of mapping the genomic coordinates of non-alignable elements between species using flanking alignable regions as anchor points. Such interpolations are most accurate if the anchor points are close, a requirement that is often not met when comparing species over large evolutionary distances, e.g. mammals and teleost fish. This is because the genomes of the members of these clades largely diverged since the last common ancestor 435 **mya**, resulting in a low density of alignable regions.

However, a particular species pair from the two clades, e.g. mouse and zebrafish, might share a partially different set of genomic elements than other mammal - teleost species pairs. In addition, due to variable evolutionary pressure and thus varying local mutation rates, some non-alignable elements between zebrafish and mouse might both be alignable to a third species that underwent different mutation events. This can be understood as follows: Let us assume that for two sequences to be alignable requires at least 70% sequence identity. Let us further assume that two sequences A and B originating from a common ancestral sequence acquired random mutations over time such that they now differ in 20% compared to the ancestral sequence. If the acquired random mutations of the two sequences are mutually exclusive, they would now only share 60% sequence identity and not be alignable. However, both sequences would still be alignable to the ancestral sequence. If a third sequence C originating from the same ancestral sequence acquired mutations at a much lower rate, it is possible that A and B are both alignable to C. In that scenario, C would act as a *bridging* sequence for A and B.

Even if a genomic element is not indirectly alignable through a bridging species, its anchor points from the reference to the bridging species and from the bridging to the target species might be arranged more closely than the anchor points from the reference directly to the target. Using a bridging species not only for indirect alignments but also for the overall increase of anchor point density is therefore expected to increase the mapping accuracy for genomic coordinates by interpolation between anchors. Moreover, using the optimal bridging species for a given genomic location and furthermore a set of multiple bridging species led to the conception of the **Independent Point Projection** (**IPP**) method presented in **Subsection 5.2.1**.

IPP implements **Dijkstra's Shortest Path Algorithm** for finding the optimal set of bridging species that minimize a distance function and thus maximize projection accuracy. In **Subsection 5.3.1** I described the application of **IPP** to map predicted enhancers in the developing limb bud of mouse embryos to genomic coordinates of chicken. For that, I chose a set of bridging species from within respective monophyletic groups encompassing mouse (Mammalia) and chicken (Sauria) as well as from a set of outgroup species with special emphasis on species that have been found to evolve slowly (e.g. elephant shark).

Only 307 out of 3181 predicted enhancers (~10%) are directly alignable from mouse to chicken and thus considered sequence-conserved (class *S*). **IPP** identified an additional set of 419 enhancers (~13%, class *F*) to be indirectly alignable through an optimal set of bridging species, and these elements exhibit enhancer-like features in chicken resembling those of the sequence-conserved enhancers. In addition, *F* en-

hancers are more likely to be sequence-conserved across multiple vertebrates similar to *S* enhancers. If an enhancer is alignable between two species, it is likely also conserved in other species. These results, however, reject the reverse conclusion, and the method is capable of differentiating between enhancers that show no evidence of conservation, and enhancers that are potentially functionally conserved.

Extending the set from putative enhancers predicted by **eHMM** to the rather loose criterion of **ATAC-seq** peaks rendered 16,865 putative cis-regulatory elements of which 2608 (~ 15%) are sequence-conserved between mouse and chicken (*S*). **IPP** identified another 3023 candidates for functionally conserved elements (*F*). Again, elements in classes *S* and *F* are highly similar with respect to functional epigenomic features such as **histone modifications**, chromatin accessibility, **RNAP II** occupancy and the genomic feature of sequence conservation to other vertebrates.

Different genomic loci appear to vary substantially in the abundance of functional conservation (**Figure 5.9**). While classes *S* and *F* comprise about 50% of putative regulatory elements in the *Sox9* **GRB**, this fraction is considerably lower in the **GRB** encompassing *Fat1* (~10%).

The chosen paths through the species graph that maximize projection accuracy have variable lengths ranging from one to six species. The vast majority of all paths and especially of shorter paths (1-2 bridging species) lead through mammalian species. This is likely a consequence of the mouse's relatively high rate of somatic mutations [270], which results in its genome having accumulated more changes than the average mammal since their common ancestor with chicken. However, some of those changes are shared among mammals, making them the ideal relay for genomic projections to other clades. In fact, a perfect bridging species for mouse has an ideal balance between being closely related to mouse and having a low mutation rate itself, so that it maximally resembles the mouse as well as the chicken. Indeed, smaller evolutionary distances from a species to either mouse or chicken make it more likely to appear in the bridging species path as demonstrated by human and xenopus within the mammalian clade and the outgroup, respectively. The exception is the Sauria clade with alligator occurring more frequently than the two birds ostrich and penguin despite the former having diverged from birds more than 100 **myr** before their speciation. This can be explained by either mutation rates (e.g. low in alligator, high in ostrich and penguin), or, more likely, by technical reasons. For example, low quality genome assemblies negatively affect alignability and highly fragmented scaffolds may often not fulfill synteny requirements and be excluded during **IPP**'s test for collinearity.

In a comprehensive application on 152 entire **GRBs** between zebrafish and mouse (**Subsection 5.3.3**), **IPP** increased the anchor point density more than fivefold compared to direct alignments. Consequently, **IPP** produces accurate genomic projections with an average distance of 500 and 600 **bps** between projected points to their closest anchor in zebrafish and mouse, respectively. This contrasts with these distances averaging at 4.8 and 8.7 **kbp** using direct alignments. According to this measure, **IPP**'s multi-species approach increases projection accuracy more than tenfold.

Altogether, **IPP** provides a valuable means to accurately project genomic regions across species with large evolutionary distances. The findings presented in this Chapter suggest that large fractions of allegedly non-conserved regulatory elements might effectively be conserved in function. I propose functional conservation beyond the conservation of sequence to be a widespread phenomenon. I encourage further studies to use these regions as candidates for experimental reporter assays in order to test them for regulatory activity in chicken and thus gain a broader understanding about regulatory landscapes and the evolutionary dynamics that shaped them. Moreover, the present results encourage large scale investigations across all **GRBs** in order to examine the degree to which regulatory elements in different loci are conserved, turned over or species-specific and the potential consequences this might entail for matters like chromatin topology and gene expression. Including multiple developmental stages and tissues would enable obtaining a more complete picture about time- and cell-type-dependent patterns of conservation. Further expanding the analysis to other species would allow tracking trajectories of enhancer evolution. Dissecting enhancers in order to discover their functional core presents another promising direction of future studies. For example, investigating enhancers where function was lost despite the sequence being largely conserved could contribute towards deciphering the relevant parts of an enhancer that encode function.

In **Subsection 5.3.4** I assessed the conservation of the **histone modification H3K27me3** between zebrafish and mouse **GRBs** and found occurrences of epigenomic conservation in absence of conserved sequence. I divided the **GRBs** based on local enrichment and conservation of **H3K27me3** into the three subdomains α (mutually enriched), β (mutually depleted) and δ (differentially enriched in zebrafish or mouse). Putative regulatory elements in α are more likely to be sequence-conserved to other vertebrates, however, approximately two thirds of them do not directly align to mouse and roughly 15% are specific to zebrafish (phastCons score < 0.05), demonstrating that epigenetic conservation is not necessarily a product of conserved sequence but can be maintained in absence thereof. Elements in α have a slight tendency to be more proximal to domain boundaries, a property that has been found to correlate with enhancers that contact multiple targets during loop extrusion. However, the ef-

fect size is rather small and does not allow the conclusion that enhancers in α are categorically different from enhancers in other subdomains in terms of topological properties. It would be interesting to acquire high resolution chromosome conformation capture data to further investigate differences in chromatin topology between the subdomains.

In order to study enhancer equivalence I developed a method called **Syntenic Anchor Point Propagation** (**SAPP**). Similar to **IPP**, this method relies on pairwise anchor points between multiple species. However, it propagates the anchors through the species graph with the goal of minimizing the resulting anchor span in the target species while maintaining synteny rather than independently projecting genomic point coordinates. **SAPP** successfully increases the projection accuracy by decreasing target anchor spans more than three fold compared to using direct alignments. It thus provides the means to narrow the search field for the analysis of enhancer equivalence.

As expected, putative regulatory elements are more likely to be homotypic when they are directly alignable between zebrafish and mouse. However, the motif-match distribution of non-alignable elements resembles that of the alignable elements only in the case of the α subdomain, where conservation of motif composition and epigenomic features correlate. This suggests that sequences of regulatory elements can diverge to a large extent - as far as losing alignability - without losing their core function, the ability to bind a particular set of **TFs**. The expected degree of freedom for sequences to change while maintaining function is possibly high as the order of the **TFBSs** for those factors potentially does not always matter.

It is important to note that **TF** motifs are not necessarily a direct indicator of function and that the resulting sample size in this analysis was rather small. Blank generalizations of these findings are therefore not possible. Instead, I propose future studies with refined methods for a more sophisticated comparison of sequence function beyond alignability. Moreover, the comparison between mouse and zebrafish is inherently difficult because of the large evolutionary distance and zebrafish's additional whole genome duplication event. Comparative studies of more closely related species could potentially help increasing sample sizes.

Together, the efforts presented in this Chapter aim to contribute towards understanding evolutionary aspects of gene regulatory logic. I presented the means for the identification of candidates with conserved function despite diverged sequences. Future methods will highly profit from experimental validations of those candidates to learn more about the sequence features of functional orthologs. My work regarding iden-

tifying functionally equivalent enhancers is only scratching the surface, touching on the most prominent examples. As a scientific community, we need to learn more about how to integrate sequence information at enhancers and allocate it to their function.

6 DISCUSSION AND CONCLUSION

In this chapter I will discuss the presented work on the prediction of enhancers and the analysis of their evolutionary conservation. I will relate my findings to today's state of knowledge and review the limitations of the presented methods and how they could be improved. Finally, I will provide a perspective on possible future directions for research and lastly conclude the thesis.

6.1 DISCUSSION

In this thesis I examined the descriptive as well as the evolutionary properties of cis-regulatory elements in the genomes of vertebrate species. Thorough inspection of the former revealed that enhancers can be described by a plethora of features and that no individual feature alone as well as no single combination of features by itself would identify all enhancers. Unfortunately, there is no gold standard set of enhancers that would qualify as a comprehensive validation set for computational approaches to identify enhancers, and even the mere number of enhancers in the human genome is subject to debate, with speculations ranging over several orders of magnitude. My supervisors and I thus put the focus on designing a method that specializes on particular aspects of enhancer prediction that had not been sufficiently addressed before.

First, I learned that many well-established methods such as the unsupervised segmentation method chromHMM suffer from a low specificity for the sake of high sensitivity and usually predict in the order of hundreds of thousands enhancers per given cell type. Second, many methods are blind to the heterogeneous molecular composition of enhancers and consider their quantitative features averaged over a predefined range, despite the established perception that enhancers comprise a central stretch of accessible **DNA** flanked by nucleosomal **DNA** wrapped around histone proteins that carry post-translational modifications that can be quantified by experimental techniques such as **ChIP-seq**. Third, some methods were found to require experimental data from laborious or expensive experiments and others performed well in certain scenarios, but lacked generalizability over vast ranges of tissues and developmental stages.

These observations led to the conception of **enhancer Hidden Markov Model** (**eHMM**), which uses a constricted and custom-tailored supervised **hidden Markov model** (**HMM**) to predict enhancers and promoters. **eHMM** requires minimal input data

in the form of four features that represent chromatin accessibility (**ATAC-seq**), transcriptional activity (**H3K27ac**) and the possibility to distinguish enhancers from promoters (**mono- and trimethylation of histone 3, lysine 4 (H3K4me1/3)**). These are widely-used features that are abundantly available online for many species, tissues and developmental stages, as well as cheaply producible in standard laboratory facilities. **eHMM** is fast and computationally cheap, especially compared to deep learning methods. It produces high-confidence predictions that are highly specific at the expense of low sensitivity, typically in the order of a few thousand per sample. **eHMM** provides a pre-trained model that can readily be applied to any cell type and uses quantile normalization to adjust read count distributions across samples. **eHMM** outperforms the state-of-the-art method REPTILE in terms of robustness across samples and resistance to overfitting, especially when testing on data from different sources than the training data. The spatial accuracy of enhancer predictions by **eHMM** is remarkably high thanks to the distinction between accessible and nucleosomal chromatin. This consideration of molecular structure is the key feature of **eHMM** that allows it to keep a low false positive rate.

Naturally, the strong focus on particular aspects of enhancer prediction implicated certain trade-offs. The low number of predicted enhancers can be a problem when aiming for statistical analyses for which larger sample sizes are beneficial. Also, the method is fixed on four features and does not include the option to include additional features upon availability. This certainly presents a promising potential for future development of the method. Moreover, expanding the method to call different types of enhancers, e.g. transcriptionally active enhancers based on **CAGE**-tags or evolutionarily conserved enhancers exhibiting active chromatin marks or characteristic **DNA** methylation patterns would broaden the scope and allow for more diversity among the predicted elements. Currently, the field of molecular genomics experiences a shift from bulk data from whole cell populations towards single-cell genomics. This entails new challenges for enhancer prediction methods such as sparse data and it will be interesting to see how future methods deal with those.

In the early years of enhancer prediction particular attention was paid to sequence conservation as a major feature of enhancers. Of course, a conserved sequence implies function and if the conserved region is located distal to a **TSS**, it is likely an enhancer. Since then, studies have shown that enhancers emerge, change, turn over and disappear in a highly dynamic fashion over time and that only a subset of all enhancers is conserved. In addition, recent reports about individual occurrences of functional conservation in absence of sequence conservation have added another perspective on the evolutionary dynamics of enhancers. According to those, an enhancer can appear to be specific to a particular species or clade because its sequence diverged to an extent

rendering it unalignable to other species with large evolutionary distances, when in fact it may have retained function nonetheless. I set out to identify such events of functional conservation and developed two methods called **Independent Point Projection** (**IPP**) and **Syntenic Anchor Point Propagation** (**SAPP**).

IPP projects genomic coordinates between two species by interpolation of the relative location between two alignable anchor points. Such alignable anchor points are sparsely distributed if the evolutionary distance between the compared species is large, and projections by interpolation with large distances to the anchors are inaccurate. **IPP** therefore uses anchor points to intermediate species in order to increase the anchor point density and subsequently projection accuracy. It does so for multiple sequential species by finding the optimal combination of bridging species through a multi-species graph using **Dijkstra's Shortest Path Algorithm**, with bridging species for which the anchors are close to the projected element being favored.

Applying **IPP** on 3181 predicted enhancers by **eHMM**, I identified 419 candidates for functional conservation in chicken in addition to the 307 predicted enhancers with conserved sequence. Extending the analysis to 16,870 **ATAC-seq** peaks, a more loose definition of regulatory elements including promoters, **IPP** predicts 3023 of those to be potentially functionally conserved. Similarly to the predicted enhancers, this fraction is again larger than the 2608 sequence-conserved elements. In chicken, the projected locations of putative functionally conserved elements exhibit epigenomic properties associated with enhancers and promoters and tend to be conserved to a larger number of vertebrates than the non-conserved elements. Overall, epigenomic features of identified candidate elements highly resemble those of the sequence-conserved elements, suggesting conserved function in absence of sequence conservation. Some of these candidates may look like they are functionally conserved on the basis of being alignable through bridging species when they are truly not. Of course, this is true for sequence-conserved elements as well. Just because we can align a sequence does not mean that function is maintained. It is therefore crucial to validate the extent to which putatively conserved elements are functional in experimental reporter assays. It will be interesting to compare the fractions of effectively functionally conserved elements among the putative candidates that are directly sequence-conserved or only indirectly as identified via bridging species.

Conservation of regulatory elements appears to strongly differ between individual loci, with the fraction of directly or indirectly conserved putative regulatory elements ranging from 10% (*Fat1*) to 50% (*Sox9*) among the limb-specific loci assessed in **Subsection 5.3.1**. The extent to which gene regulation is conserved thus seems to vary from locus to locus. Future research addressing the reconstruction of evolutionary

trajectories of enhancers in specific loci taking into account enhancer turnover will be necessary to elucidate the origins and emergence of gene regulatory logic.

A comprehensive quantitative evaluation of **IPP**'s projection quality demonstrated that using the multi-species approach of **IPP** greatly elevated anchor density and thus decreased the average distance of a projected genomic location to its anchor points in all the reference, bridging and target species compared to using only direct anchor points. With that, I applied it in an extensive approach to map the signal of **H3K27me3** from 152 **GRBs** from mouse onto zebrafish genomic coordinates and assessed signal overlap. Roughly 10 % of zebrafish chromatin in **GRBs** is enriched for **H3K27me3**. According to the results from mapping the mouse signal onto zebrafish coordinates, approximately 40% of that is shared with mouse. However, many putative regulatory elements within those regions of shared epigenomics are not conserved in sequence. I hypothesized that regions of shared epigenomics could represent topological subdomains due to the fact that **H3K27me3** is a readout of current or past Polycomb activity which in turn is known for its contribution to the stabilization of chromatin loops and establishing chromatin domains. According to that, I identified three classes of subdomains which are either mutually enriched or mutually depleted for **H3K27me3** in zebrafish and mouse, or differentially enriched in zebrafish only. Regulatory elements in the mutually enriched subdomain are more likely to be located at domain boundaries. However, the effect size of this result is very small and whether the identified subdomains actually represent topological units remains elusive and may be subject to future research.

In addition to the quantitative improvement of projection accuracy, I demonstrated **IPP**'s capability of identifying individual occurrences of putative functionally conserved elements. As demonstrated in **Subsection 5.3.2**, **IPP** correctly projected the genomic coordinates of a putative regulatory element in mouse onto a region in chicken that overlaps epigenomic features that suggest enhancer function. It did so for a genomic region vastly devoid of directly alignable sequences, exemplifying the benefit of individually optimizing bridging species selection. These findings encourage the conceptualization of a future project involving the systematic search for such cases of putative functional conservation followed by experimental validation in order to draft a catalog of functionally conserved elements beyond sequence conservation.

To test functional equivalence of enhancers under a different perspective, I developed the method **SAPP**. Instead of projecting independent point coordinates, **SAPP** takes the anchor points of a given genomic location and propagates them through the bridging species until the target species is reached, selecting the path which minimizes the target anchor span. With that, the resulting anchor span is expected to contain a po-

tentially equivalent enhancer given that synteny is conserved. Minimizing the target anchor span thus allows narrowing the search field as much as possible in order to exclude non-equivalent enhancers. This resulted in the three fold reduction of anchor spans, providing a good foundation for studying enhancer equivalence. For that, I assessed the **TFBS** motif composition of elements and quantified enhancer similarity based on the number of motif matches. On average, directly alignable elements share more motifs with their aligned counterpart than with randomly assigned elements. When testing the non-alignable elements, only those in the epigenomically conserved subdomains in zebrafish were more likely to share motifs with the mouse elements that are located within the anchors than randomly assigned elements across anchor boundaries. These findings suggest that enhancer equivalence is more likely when epigenomic features are conserved, even when the sequence is not. It is possible that regulatory elements within epigenomically conserved subdomains are more likely to be functionally conserved which in turn might contribute to the maintenance of the epigenomic, and ultimately topological properties. The limitation of this analysis certainly is the small sample size. As a consequence of the large evolutionary distance between zebrafish and mouse, many enhancers are species-specific, and out of the non-alignable elements located within the epigenomically conserved subdomains, only very few had at least one counterpart element within the anchor boundaries in mouse. These results therefore need to be regarded with a certain level of caution.

6.2 CONCLUSION

In this thesis I presented three methods called **eHMM**, **IPP** and **SAPP**. **eHMM** predicts enhancers based on epigenomic features with a prediction performance equal to or higher than state-of-the-art methods. **eHMM** is robustly applicable to multiple tissues and developmental stages and its predictions exhibit a remarkably high spatial accuracy. I demonstrated that **HMMs** can be used in a highly modular and constricted way by imposing constraints on transition parameters as well as supervise their training on multiple training sets. **IPP** and **SAPP** project non-alignable elements between species with large evolutionary distances by increasing anchor point density using multiple pairwise species comparisons. Both methods substantially increase projection accuracy compared to using sparsely distributed direct anchors and help identify candidate enhancers for functional conservation despite diverged sequences.

Together, these efforts have contributed to the understanding of enhancers' epigenomic properties as well as their evolutionary dynamics, and have opened up new questions for future research.

APPENDIX A

A.1 MATHEMATICAL DERIVATIONS

A.1.1 OPTIMIZING THE Q-FUNCTION SUBJECT TO CONSTRAINTS USING LAGRANGE MULTIPLIERS

According to **Equation 3.14**, the Q-function is given by

$$Q\left(\theta' \mid \theta\right) = \sum_{x \in Z^L} p(x) \left(\log \pi_{x_1} + \sum_{l=1}^{L-1} \log a_{x_l x_{l+1}} + \sum_{l=1}^{L} \log b_{x_l}(y_l) \right).$$

Finding the optimal parameters of the Q-function is subject to the following constraints:

$$\sum_{i=1}^{N} \pi_i = 1, \quad \sum_{j=1}^{N} a_{ij} = 1, \quad \sum_{k \in K} b_i(k) = 1$$

with K denoting the set of all possible observables. The constrained optimization problem can then be formulated as the Lagrange function:

$$\begin{aligned} \mathcal{L}(\theta, \lambda) = & \sum_{x \in Z^L} p(x) \left(\log \pi_{x_1} + \sum_{l=1}^{L-1} \log a_{x_l x_{l+1}} + \sum_{l=1}^{L} \log b_{x_l}(y_l) \right) \\ & - \lambda_1 \left(\sum_{i=1}^{N} \pi_i - 1 \right) - \lambda_2 \left(\sum_{j=1}^{N} a_{ij} - 1 \right) - \lambda_3 \left(\sum_{k \in K} b_i(k) - 1 \right) \end{aligned}$$

I will show the optimization of each parameter by calculating the roots of the respective partial derivatives of the Lagrange function in the following subsections.

A.1.2 OPTIMIZING THE INITIAL PROBABILITIES

$$\nabla_{\pi_i, \lambda} \mathcal{L}(\pi, \lambda) = 0$$

$$\frac{\partial \mathcal{L}(\pi, \lambda)}{\partial \pi_i} = \frac{\partial \left(\sum_{i=1}^{N} \gamma_i(1) \log \pi_i - \lambda(\sum_i \pi_i - 1) \right)}{\partial \pi_i} = 0$$

$$0 = \frac{\gamma_i(1)}{\pi_i} - \lambda$$

$$\pi_i = \frac{\gamma_i(1)}{\lambda}$$

We can now reformulate the constraint:

$$1 = \sum_i \pi_i = \frac{\sum_i \gamma_i(1)}{\lambda}$$

$$\lambda = \sum_i \gamma_i(1) = 1$$

The previously determined π_i then becomes

$$\pi_i = \gamma_i(1)$$

A.1.3 OPTIMIZING THE TRANSITION PROBABILITIES

$$\nabla_{a_{ij}, \lambda} \mathcal{L}(a, \lambda) = 0$$

$$\frac{\partial \mathcal{L}(a, \lambda)}{\partial a_{ij}} = \frac{\partial \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{l=1}^{L-1} \xi_{ij}(l) \log a_{ij} - \lambda(\sum_j a_{ij} - 1) \right)}{\partial a_{ij}} = 0$$

$$0 = \frac{\sum_{l=1}^{L-1} \xi_{ij}(l)}{a_{ij}} - \lambda$$

$$a_{ij} = \frac{\sum_{l=1}^{L-1} \xi_{ij}(l)}{\lambda}$$

We can now reformulate the constraint:

$$1 = \sum_j a_{ij} \qquad = \frac{\sum_j \sum_{l=1}^{L-1} \xi_{ij}(l)}{\lambda}$$

$$\lambda = \sum_j \sum_{l=1}^{L-1} \xi_{ij}(l) \quad = \sum_{l=1}^{L-1} \gamma_i(l)$$

The previously determined a_{ij} then becomes

$$a_{ij} = \frac{\sum_{l=1}^{L-1} \xi_{ij}(l)}{\sum_{l=1}^{L-1} \gamma_i(l)}$$

A.1.4 OPTIMIZING THE EMISSION PROBABILITIES

$$\nabla_{b_i(k),\lambda} \mathcal{L}(b, \lambda) = 0$$

$$\frac{\partial \mathcal{L}(b,\lambda)}{\partial b_i(k)} = \frac{\partial \left(\sum_{i=1}^{N} \sum_{l=1}^{L} \gamma_i(l) \log b_i(k) - \lambda(\sum_k b_i(k) - 1) \right)}{\partial b_i(k)} \quad = 0$$

$$0 = \frac{\sum_{l=1}^{L} \gamma_i(l)}{b_i(k)} - \lambda$$

$$b_i(k) = \frac{\sum_{l=1}^{L} \gamma_i(l)}{\lambda}$$

We can now reformulate the constraint:

$$1 = \sum_{k \in K} b_i(k) \qquad = \frac{\sum_{k \in K} \sum_{l=1}^{L} \gamma_i(l)}{\lambda}$$

$$\lambda = \sum_{k \in K} \sum_{l=1}^{L} \gamma_i(l) \quad = n_K \sum_{l=1}^{L} \gamma_i(l)$$

The previously determined $b_i(k)$ then becomes

$$b_i(k) = \frac{\sum_{l=1}^{L} \gamma_i(l)\ \mathbb{1}(k = y_l)}{\sum_{l=1}^{L} \gamma_i(l)}$$

A.1.5 OPTIMIZING THE PARAMETERS OF THE LOG-NORMAL DISTRIBUTION MODELING THE EMISSION PROBABILITIES

$$\begin{aligned}
\frac{\partial Q(\theta' \mid \theta)}{\partial \mu_i} &= \frac{\partial \left(\sum_{l=1}^{L} \gamma_i(l) \log f_{LN}(y_l; \mu_i, \sigma_i) \right)}{\partial \mu_i} \\
&= \frac{\partial \left(\sum_{l=1}^{L} \gamma_i(l) \left(-\ln\left(y_l \sigma_i \sqrt{2\pi}\right) - \frac{(\ln y_l - \mu_i)^2}{2\sigma_i^2} \right) \right)}{\partial \mu_i} \\
&= \sum_{l=1}^{L} \gamma_i(l)\ \frac{\ln y_l - \mu_i}{\sigma_i^2} \qquad = 0 \\
\mu_i \sum_{l=1}^{L} \gamma_i(l) &= \sum_{l=1}^{L} \gamma_i(l) \ln y_l \\
\mu_i &= \frac{\sum_{l=1}^{L} \gamma_i(l) \ln y_l}{\sum_{l=1}^{L} \gamma_i(l)}
\end{aligned}$$

$$\frac{\partial Q(\theta' \mid \theta)}{\partial \sigma_i} = \frac{\partial \left(\sum_{l=1}^{L} \gamma_i(l) \log f_{LN}(y_l; \mu_i, \sigma_i) \right)}{\partial \sigma_i}$$

$$= \frac{\partial \left(\sum_{l=1}^{L} \gamma_i(l) \left(-\ln \left(y_l \sigma_i \sqrt{2\pi} \right) - \frac{(\ln y_l - \mu_i)^2}{2\sigma_i^2} \right) \right)}{\partial \sigma_i}$$

$$= \sum_{l=1}^{L} \gamma_i(l) \left(-\frac{1}{\sigma} + \frac{(\ln y_l - \mu_i)^2}{\sigma^3} \right)$$

$$= \frac{1}{\sigma^3} \sum_{l=1}^{L} \gamma_i(l) \left(-\sigma_i^2 + (\ln y_l - \mu_i)^2 \right) \qquad = 0$$

$$\sigma_i^2 \sum_{l=1}^{L} \gamma_i(l) = \sum_{l=1}^{L} \gamma_i(l) \, (\ln y_l - \mu_i)^2$$

$$\sigma_i = \sqrt{\frac{\sum_{l=1}^{L} \gamma_i(l) \, (\ln y_l - \mu_i)^2}{\sum_{l=1}^{L} \gamma_i(l)}}$$

A.2 DATA SOURCES

A.2.1 FUNCTIONAL GENOMICS DATA FOR ENHANCER PREDICTION

Table 2: Data sources. Accession numbers containing GSE were obtained from **NCBI GEO** [271–274], those starting with ENC from **ENCODE** [210].

Cell type	Experiment	Target	Accession	Format
ESC E14	ATAC-seq	-	GSE120376	fastq
	ChIP-seq	H3K27ac	GSE120376	fastq
		H3K4me1	GSE120376	fastq
		H3K4me3	GSE120376	fastq
		Nanog	GSE11431	fastq
		Oct4	GSE11431	fastq
		Sox2	GSE11431	fastq
		CTCF	GSE29184	fastq
		p300	GSE29184	fastq
		Pol II	GSE29184	fastq
	MeDIP-seq	-	GSE3859	fastq
liver E12.5	ATAC-seq	-	ENCSR302LIV	bam
	ChIP-seq	H3K27ac	ENCSR136GMT	bam
		H3K4me1	ENCSR770OXU	bam
		H3K4me3	ENCSR471SJG	bam
liver E14.5	ATAC-seq	-	ENCSR032HKE	fastq
	ChIP-seq	H3K27ac	ENCSR075SNV	bam
		H3K4me1	ENCSR234ISO	bam
		H3K4me3	ENCSR433ESG	bam
lung E14.5	ATAC-seq	-	ENCSR335VJW	fastq
	ChIP-seq	H3K27ac	ENCSR452WYC	bam
		H3K4me1	ENCSR825OWH	bam
		H3K4me3	ENCSR839WFP	bam
lung E16.5	ATAC-seq	-	ENCSR627OCR	fastq
	ChIP-seq	H3K27ac	ENCSR140UEX	bam
		H3K4me1	ENCSR387YSD	bam
		H3K4me3	ENCSR295PFM	bam

BIBLIOGRAPHY

[1] Tobias Zehnder, Philipp Benner, and Martin Vingron. "Predicting enhancers in mammalian genomes using supervised hidden Markov models." *BMC Bioinformatics* 20.1 (Mar. 2019), p. 157. ISSN: 1471-2105. DOI: 10.1186/s12859-019-2708-6.

[2] F H Crick. "On protein synthesis." *Symposia of the Society for Experimental Biology* 12 (1958), pp. 138–163.

[3] Valerie A Schneider et al. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly." *Genome Research* 27.5 (Apr. 2017), pp. 849–864. DOI: 10.1101/gr.213611.116.

[4] E S Lander et al. "Initial sequencing and analysis of the human genome." *Nature* 409.6822 (Feb. 2001), pp. 860–921. DOI: 10.1038/35057062.

[5] J C Venter et al. "The sequence of the human genome." *Science* 291.5507 (Feb. 2001), pp. 1304–1351. DOI: 10.1126/science.1058040.

[6] Mihaela Pertea and Steven L Salzberg. "Between a chicken and a grape: estimating the number of human genes." *Genome Biology* 11.5 (May 2010), p. 206. DOI: 10.1186/gb-2010-11-5-206.

[7] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L Tress. "Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes." *Human Molecular Genetics* 23.22 (Nov. 2014), pp. 5866–5878. DOI: 10.1093/hmg/ddu309.

[8] Jill Cheng et al. "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution." *Science* 308.5725 (May 2005), pp. 1149–1154. ISSN: 1095-9203. DOI: 10.1126/science.1108625.

[9] M Behfar Ardehali and John T Lis. "Tracking rates of transcription and splicing in vivo." *Nature Structural & Molecular Biology* 16.11 (Nov. 2009), pp. 1123–1124. DOI: 10.1038/nsmb1109-1123.

[10] C O Pabo and R T Sauer. "Transcription factors: structural families and principles of DNA recognition." *Annual Review of Biochemistry* 61 (1992), pp. 1053–1095. DOI: 10.1146/annurev.bi.61.070192.005201.

[11] Matthew T Weirauch and T R Hughes. "A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution." *Subcellular biochemistry* 52 (2011), pp. 25–73. DOI: 10.1007/978-90-481-9069-0_3.

[12] M Madan Babu, Nicholas M Luscombe, L Aravind, Mark Gerstein, and Sarah A Teichmann. "Structure and evolution of transcriptional regulatory networks." *Current Opinion in Structural Biology* 14.3 (June 2004), pp. 283–291. DOI: 10.1016/j.sbi.2004.05.004.

[13] Andrew J Bannister and Tony Kouzarides. "Regulation of chromatin by histone modifications." *Cell Research* 21.3 (Mar. 2011), pp. 381–395. DOI: 10.1038/cr.2011.22.

[14] Mustafa Mir, Wendy Bickmore, Eileen E M Furlong, and Geeta Narlikar. "Chromatin topology, condensates and gene regulation: shifting paradigms or just a phase?" *Development* 146.19 (Sept. 2019). ISSN: 0950-1991. DOI: 10.1242/dev.182766.

[15] Elphège P Nora et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre." *Nature* 485.7398 (Apr. 2012), pp. 381–385. DOI: 10.1038/nature11049.

[16] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature* 485.7398 (Apr. 2012), pp. 376–380. DOI: 10.1038/nature11082.

[17] Adriana Gonzalez-Sandoval and Susan M Gasser. "On tads and lads: spatial control over gene expression." *Trends in Genetics* 32.8 (June 2016), pp. 485–495. DOI: 10.1016/j.tig.2016.05.004.

[18] Laura A Lettice et al. "Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly." *Proceedings of the National Academy of Sciences of the United States of America* 99.11 (May 2002), pp. 7548–7553. DOI: 10.1073/pnas.112212199.

[19] J Banerji, S Rusconi, and W Schaffner. "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." *Cell* 27.2 Pt 1 (Dec. 1981), pp. 299–308. DOI: 10.1016/0092-8674(81)90413-x.

[20] M Ptashne. "Regulation of transcription: from lambda to eukaryotes." *Trends in Biochemical Sciences* 30.6 (June 2005), pp. 275–279. DOI: 10.1016/j.tibs.2005.04.003.

[21] Len A Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A Nobrega, and Gill Bejerano. "Enhancers: five essential questions." *Nature Reviews. Genetics* 14.4 (2013), pp. 288–295. DOI: 10.1038/nrg3458.

[22] Walter Schaffner. "Enhancers, enhancers - from their discovery to today's universe of transcription enhancers." *Biological Chemistry* 396.4 (Apr. 2015), pp. 311–327. DOI: 10.1515/hsz-2014-0303.

[23] Benjamin L Allen and Dylan J Taatjes. "The Mediator complex: a central integrator of transcription." *Nature Reviews. Molecular Cell Biology* 16.3 (Mar. 2015), pp. 155–166. DOI: 10.1038/nrm3951.

[24] Thomas M Harper and Dylan J Taatjes. "The complex structure and function of Mediator." *The Journal of Biological Chemistry* 293.36 (Sept. 2018), pp. 13778–13785. DOI: 10.1074/jbc.R117.794438.

[25] P Cramer et al. "Structure of eukaryotic RNA polymerases." *Annual review of biophysics* 37 (2008), pp. 337–352. DOI: 10.1146/annurev.biophys.37.032807.130008.

[26] M S Halfon, A Carmena, S Gisselbrecht, C M Sackerson, F Jiménez, M K Baylies, and A M Michelson. "Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors." *Cell* 103.1 (Sept. 2000), pp. 63–74. DOI: 10.1016/s0092-8674(00)00105-7.

[27] C Xu, R C Kauffmann, J Zhang, S Kladny, and R W Carthew. "Overlapping activators and repressors delimit transcriptional response to receptor tyrosine kinase signals in the Drosophila eye." *Cell* 103.1 (Sept. 2000), pp. 87–97. DOI: 10.1016/s0092-8674(00)00107-0.

[28] G V Flores, H Duan, H Yan, R Nagaraj, W Fu, Y Zou, M Noll, and U Banerjee. "Combinatorial signaling in the specification of unique cell fates." *Cell* 103.1 (Sept. 2000), pp. 75–85. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(00)00106-9.

[29] A Ghazi and K V VijayRaghavan. "Developmental biology. Control by combinatorial codes." *Nature* 408.6811 (Nov. 2000), pp. 419–420. DOI: 10.1038/35044174.

[30] D P Bazett-Jones, B Leblanc, M Herfort, and T Moss. "Short-range DNA looping by the Xenopus HMG-box transcription factor, xUBF." *Science* 264.5162 (May 1994), pp. 1134–1137. DOI: 10.1126/science.8178172.

[31] François Spitz and Eileen E M Furlong. "Transcription factors: from enhancer binding to developmental control." *Nature Reviews. Genetics* 13.9 (Sept. 2012), pp. 613–626. DOI: 10.1038/nrg3207.

[32] D Thanos and T Maniatis. "Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome." *Cell* 83.7 (Dec. 1995), pp. 1091–1100. DOI: 10.1016/0092-8674(95)90136-1.

[33] M Merika and D Thanos. "Enhanceosomes." *Current Opinion in Genetics & Development* 11.2 (Apr. 2001), pp. 205–208. DOI: 10.1016/S0959-{437X}(00)00180-5.

[34] Daniel Panne, Tom Maniatis, and Stephen C Harrison. "An atomic model of the interferon-beta enhanceosome." *Cell* 129.6 (June 2007), pp. 1111–1123. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.05.019.

[35] Meghana M Kulkarni and David N Arnosti. "Information display by transcriptional enhancers." *Development* 130.26 (Dec. 2003), pp. 6569–6575. ISSN: 0950-1991. DOI: 10.1242/dev.00890.

[36] Guillaume Junion, Mikhail Spivakov, Charles Girardot, Martina Braun, E Hilary Gustafson, Ewan Birney, and Eileen E M Furlong. "A transcription factor collective defines cardiac cell fate and reflects lineage history." *Cell* 148.3 (Feb. 2012), pp. 473–486. DOI: 10.1016/j.cell.2012.01.030.

[37] Ziga Avsec et al. "Deep learning at base-resolution reveals motif syntax of the cis-regulatory code." *BioRxiv* (Aug. 2019). DOI: 10.1101/737981.

[38] Joanna A Miller and Jonathan Widom. "Collaborative competition mechanism for gene activation in vivo." *Molecular and Cellular Biology* 23.5 (Mar. 2003), pp. 1623–1632. ISSN: 0270-7306. DOI: 10.1128/mcb.23.5.1623-1632.2003.

[39] Leonid A Mirny. "Nucleosome-mediated cooperativity between transcription factors." *Proceedings of the National Academy of Sciences of the United States of America* 107.52 (Dec. 2010), pp. 22534–22539. DOI: 10.1073/pnas.0913805107.

[40] Hannah K Long, Sara L Prescott, and Joanna Wysocka. "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution." *Cell* 167.5 (Nov. 2016), pp. 1170–1187. ISSN: 00928674. DOI: 10.1016/j.cell.2016.09.018.

[41] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. "Absence of a simple code: how transcription factors read the genome." *Trends in Biochemical Sciences* 39.9 (Sept. 2014), pp. 381–399. DOI: 10.1016/j.tibs.2014.07.002.

[42] Gregory E Crawford, Sean Davis, Peter C Scacheri, Gabriel Renaud, Mohamad J Halawi, Michael R Erdos, Roland Green, Paul S Meltzer, Tyra G Wolfsberg, and Francis S Collins. "DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays." *Nature Methods* 3.7 (July 2006), pp. 503–509. DOI: 10.1038/nmeth888.

[43] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. "High-resolution mapping and characterization of open chromatin across the genome." *Cell* 132.2 (Jan. 2008), pp. 311–322. DOI: 10.1016/j.cell.2007.12.014.

[44] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature Methods* 10.12 (Dec. 2013), pp. 1213–1218. DOI: 10.1038/nmeth.2688.

[45] B D Strahl and C D Allis. "The language of covalent histone modifications." *Nature* 403.6765 (Jan. 2000), pp. 41–45. DOI: 10.1038/47412.

[46] B M Turner. "Histone acetylation and an epigenetic code." *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology* 22.9 (Sept. 2000), pp. 836–845. DOI: 10.1002/1521-1878(200009)22:9\textless836::{AID}-{BIES9\textgreater3}.0.{CO};2-X.

[47] Nathaniel D Heintzman et al. "Histone modifications at human enhancers reflect global cell-type-specific gene expression." *Nature* 459.7243 (May 2009), pp. 108–112. ISSN: 1476-4687. DOI: 10.1038/nature07829.

[48] Nevan J Krogan et al. "Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II." *Molecular and Cellular Biology* 23.12 (June 2003), pp. 4207–4218. ISSN: 0270-7306. DOI: 10.1128/mcb.23.12.4207-4218.2003.

[49] Feng Tie, Rakhee Banerjee, Carl A Stratton, Jayashree Prasad-Sinha, Vincent Stepanik, Andrei Zlobin, Manuel O Diaz, Peter C Scacheri, and Peter J Harte. "CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing." *Development* 136.18 (Sept. 2009), pp. 3131–3141. DOI: 10.1242/dev.037127.

[50] Bing Zhang, Daniel S Day, Joshua W Ho, Lingyun Song, Jingjing Cao, Danos Christodoulou, Jonathan G Seidman, Gregory E Crawford, Peter J Park, and William T Pu. "A dynamic H3K27ac signature identifies VEGFA-stimulated endothelial enhancers and requires EP300 activity." *Genome Research* 23.6 (June 2013), pp. 917–927. DOI: 10.1101/gr.149674.112.

[51] Tatsuya Nakamura, Toshiki Mori, Shinichiro Tada, Wladyslaw Krajewski, Tanya Rozovskaia, Richard Wassell, Garrett Dubois, Alexander Mazo, Carlo M Croce, and Eli Canaani. "ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation." *Molecular Cell* 10.5 (Nov. 2002), pp. 1119–1128. DOI: 10.1016/s1097-2765(02)00740-2.

[52] Yurii Sedkov, Elizabeth Cho, Svetlana Petruk, Lucy Cherbas, Sheryl T Smith, Richard S Jones, Peter Cherbas, Eli Canaani, James B Jaynes, and Alexander Mazo. "Methylation at lysine 4 of histone H3 in ecdysone-dependent development of Drosophila." *Nature* 426.6962 (Nov. 2003), pp. 78–83. ISSN: 1476-4687. DOI: 10.1038/nature02080.

[53] Nathaniel D Heintzman et al. "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." *Nature Genetics* 39.3 (Mar. 2007), pp. 311–318. ISSN: 1061-4036. DOI: 10.1038/ng1966.

[54] Zhibin Wang et al. "Combinatorial patterns of histone acetylations and methylations in the human genome." *Nature Genetics* 40.7 (July 2008), pp. 897–903. ISSN: 1546-1718. DOI: 10.1038/ng.154.

[55] H Wang, R Cao, L Xia, H Erdjument-Bromage, C Borchers, P Tempst, and Y Zhang. "Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase." *Molecular Cell* 8.6 (Dec. 2001), pp. 1207–1217. DOI: 10.1016/s1097-2765(01)00405-1.

[56] Helena Santos-Rosa, Robert Schneider, Bradley E Bernstein, Nickoletta Karabetsou, Antonin Morillon, Christoph Weise, Stuart L Schreiber, Jane Mellor, and Tony Kouzarides. "Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin." *Molecular Cell* 12.5 (Nov. 2003), pp. 1325–1332. DOI: 10.1016/s1097-2765(03)00438-6.

[57] Andrea Local et al. "Identification of H3K4me1-associated proteins at mammalian enhancers." *Nature Genetics* 50.1 (2018), pp. 73–82. DOI: 10.1038/s41588-017-0015-6.

[58] Menno P Creyghton et al. "Histone H3K27ac separates active from poised enhancers and predicts developmental state." *Proceedings of the National Academy of Sciences of the United States of America* 107.50 (Dec. 2010), pp. 21931–21936. DOI: 10.1073/pnas.1016071107.

[59] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A Brugmann, Ryan A Flynn, and Joanna Wysocka. "A unique chromatin signature uncovers early developmental enhancers in humans." *Nature* 470.7333 (Feb. 2011), pp. 279–283. ISSN: 1476-4687. DOI: 10.1038/nature09692.

[60] Yingming Zhao and Benjamin A Garcia. "Comprehensive catalog of currently documented histone modifications." *Cold Spring Harbor Perspectives in Biology* 7.9 (Sept. 2015), a025064. DOI: 10.1101/cshperspect.a025064.

[61] A P Bird. "CpG-rich islands and the function of DNA methylation." *Nature* 321.6067 (May 1986), pp. 209–213. DOI: 10.1038/321209a0.

[62] Michael J Ziller et al. "Charting a dynamic DNA methylation landscape of the human genome." *Nature* 500.7463 (Aug. 2013), pp. 477–481. DOI: 10.1038/nature12433.

[63] M Gardiner-Garden and M Frommer. "CpG islands in vertebrate genomes." *Journal of Molecular Biology* 196.2 (July 1987), pp. 261–282. DOI: 10.1016/0022-2836(87)90689-9.

[64] Michael B Stadler et al. "DNA-binding factors shape the mouse methylome at distal regulatory regions." *Nature* 480.7378 (Dec. 2011), pp. 490–495. ISSN: 0028-08361476-4687. DOI: 10.1038/nature10716.

[65] Hongbo Liu et al. "Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes." *Nucleic Acids Research* 44.1 (Jan. 2016), pp. 75–94. DOI: 10.1093/nar/gkv1332.

[66] Michael Klutstein, Deborah Nejman, Razi Greenfield, and Howard Cedar. "DNA methylation in cancer and aging." *Cancer Research* 76.12 (June 2016), pp. 3446–3450. DOI: 10.1158/0008-5472.{CAN}-15-3278.

[67] K Wyatt McMahon, Enusha Karunasena, and Nita Ahuja. "The roles of DNA methylation in the stages of cancer." *Cancer Journal* 23.5 (2017), pp. 257–261. DOI: 10.1097/{PPO}.0000000000000279.

[68] Peter A Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond." *Nature Reviews. Genetics* 13.7 (May 2012), pp. 484–492. DOI: 10.1038/nrg3230.

[69] Axel Visel et al. "ChIP-seq accurately predicts tissue-specific activity of enhancers." *Nature* 457.7231 (Feb. 2009), pp. 854–858. ISSN: 1476-4687. DOI: 10.1038/nature07730.

[70] Tae-Kyung Kim et al. "Widespread transcription at neuronal activity-regulated enhancers." *Nature* 465.7295 (May 2010), pp. 182–187. ISSN: 1476-4687. DOI: 10.1038/nature09033.

[71] Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. "Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers." *Nature Genetics* 46.12 (Dec. 2014), pp. 1311–1320. DOI: 10.1038/ng.3142.

[72] Frederic Koch et al. "Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters." *Nature Structural & Molecular Biology* 18.8 (July 2011), pp. 956–963. DOI: 10.1038/nsmb.2085.

[73] Olga Mikhaylichenko, Vladyslav Bondarenko, Dermot Harnett, Ignacio E Schor, Matilda Males, Rebecca R Viales, and Eileen E M Furlong. "The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription." *Genes & Development* 32.1 (Jan. 2018), pp. 42–57. DOI: 10.1101/gad.308619.117.

[74] Michael T Y Lam et al. "Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription." *Nature* 498.7455 (June 2013), pp. 511–515. DOI: 10.1038/nature12209.

[75] Carlos A Melo et al. "eRNAs are required for p53-dependent enhancer activity and gene transcription." *Molecular Cell* 49.3 (Feb. 2013), pp. 524–535. DOI: 10.1016/j.molcel.2012.11.021.

[76] Minna U Kaikkonen et al. "Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription." *Molecular Cell* 51.3 (Aug. 2013), pp. 310–325. DOI: 10.1016/j.molcel.2013.07.010.

[77] Kevin Struhl. "Transcriptional noise and the fidelity of initiation by RNA polymerase II." *Nature Structural & Molecular Biology* 14.2 (Feb. 2007), pp. 103–105. ISSN: 1545-9993. DOI: 10.1038/nsmb0207-103.

[78] Robert S Young, Yatendra Kumar, Wendy A Bickmore, and Martin S Taylor. "Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers." *Genome Biology* 18.1 (Dec. 2017), p. 242. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1379-8.

[79] Toshiyuki Shiraki et al. "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." *Proceedings of the National Academy of Sciences of the United States of America* 100.26 (Dec. 2003), pp. 15776–15781. DOI: 10.1073/pnas.2136655100.

[80] Dig Bijay Mahat, Hojoong Kwak, Gregory T Booth, Iris H Jonkers, Charles G Danko, Ravi K Patel, Colin T Waters, Katie Munson, Leighton J Core, and John T Lis. "Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq)." *Nature Protocols* 11.8 (July 2016), pp. 1455–1476. DOI: 10.1038/nprot.2016.086.

[81] Rui Lopes, Reuven Agami, and Gozde Korkmaz. "GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression." *Methods in Molecular Biology* 1543 (2017), pp. 45–55. DOI: 10.1007/978-1-4939-6716-2_3.

[82] Benjamin S Scruggs, Daniel A Gilchrist, Sergei Nechaev, Ginger W Muse, Adam Burkholder, David C Fargo, and Karen Adelman. "Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin." *Molecular Cell* 58.6 (June 2015), pp. 1101–1112. DOI: 10.1016/j.molcel.2015.04.006.

[83] Telmo Henriques, Benjamin S Scruggs, Michiko O Inouye, Ginger W Muse, Lucy H Williams, Adam B Burkholder, Christopher A Lavender, David C Fargo, and Karen Adelman. "Widespread transcriptional pausing and elongation control at enhancers." *Genes & Development* 32.1 (Jan. 2018), pp. 26–41. DOI: 10.1101/gad.309351.117.

[84] Robin Andersson et al. "An atlas of active enhancers across human cell types and tissues." *Nature* 507.7493 (Mar. 2014), pp. 455–461. DOI: 10.1038/nature12787.

[85] M Kimura. "The neutral theory of molecular evolution." *Scientific American* 241.5 (Nov. 1979), 98–100, 102, 108 passim. DOI: 10.1017/{CBO9780511623486}.

[86] Gregory M Cooper and Christopher D Brown. "Qualifying the relationship between sequence conservation and molecular function." *Genome Research* 18.2 (Feb. 2008), pp. 201–205. DOI: 10.1101/gr.7205808.

[87] B Charlesworth, M T Morgan, and D Charlesworth. "The effect of deleterious mutations on neutral molecular variation." *Genetics* 134.4 (Aug. 1993), pp. 1289–1303.

[88] Seila Omer, Timothy J Harlow, and Johann Peter Gogarten. "Does sequence conservation provide evidence for biological function?" *Trends in Microbiology* 25.1 (2017), pp. 11–18. DOI: 10.1016/j.tim.2016.09.010.

[89] Gill Bejerano, Michael Pheasant, Igor Makunin, Stuart Stephen, W James Kent, John S Mattick, and David Haussler. "Ultraconserved elements in the human genome." *Science* 304.5675 (May 2004), pp. 1321–1325. ISSN: 1095-9203. DOI: 10.1126/science.1098119.

[90] Albin Sandelin, Peter Bailey, Sara Bruce, Pär G Engström, Joanna M Klos, Wyeth W Wasserman, Johan Ericson, and Boris Lenhard. "Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes." *BMC Genomics* 5.1 (Dec. 2004), p. 99. ISSN: 1471-2164. DOI: 10.1186/1471-2164-5-99.

[91] Adam Woolfe et al. "Highly conserved non-coding sequences are associated with vertebrate development." *PLoS Biology* 3.1 (Jan. 2005), e7. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0030007.

[92] Pär G Engström, Shannan J Ho Sui, Oyvind Drivenes, Thomas S Becker, and Boris Lenhard. "Genomic regulatory blocks underlie extensive microsynteny conservation in insects." *Genome Research* 17.12 (Dec. 2007), pp. 1898–1908. ISSN: 1088-9051. DOI: 10.1101/gr.6669607.

[93] Yoichiro Shibata et al. "Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection." *PLoS Genetics* 8.6 (June 2012), e1002789. DOI: 10.1371/journal.pgen.1002789.

[94] Joung-Woo Hong, David A Hendrix, and Michael S Levine. "Shadow enhancers as a source of evolutionary novelty." *Science* 321.5894 (Sept. 2008), p. 1314. DOI: 10.1126/science.1160631.

[95] M Z Ludwig, C Bergman, N H Patel, and M Kreitman. "Evidence for stabilizing selection in a eukaryotic enhancer element." *Nature* 403.6769 (Feb. 2000), pp. 564–567. ISSN: 0028-0836. DOI: 10.1038/35000615.

[96] Gizem Kalay and Patricia J Wittkopp. "Nomadic enhancers: tissue-specific cis-regulatory elements of yellow have divergent genomic positions among Drosophila species." *PLoS Genetics* 6.11 (Nov. 2010), e1001222. DOI: 10.1371/journal.pgen.1001222.

[97] Diego Villar et al. "Enhancer evolution across 20 mammalian species." *Cell* 160.3 (Jan. 2015), pp. 554–566. DOI: 10.1016/j.cell.2015.01.006.

[98] Michael Z Ludwig, Arnar Palsson, Elena Alekseeva, Casey M Bergman, Janaki Nathan, and Martin Kreitman. "Functional evolution of a cis-regulatory module." *PLoS Biology* 3.4 (Apr. 2005), e93. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0030093.

[99] Emily E Hare, Brant K Peterson, Venky N Iyer, Rudolf Meier, and Michael B Eisen. "Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation." *PLoS Genetics* 4.6 (June 2008), e1000106. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000106.

[100] Leila Taher, David M McGaughey, Samantha Maragh, Ivy Aneas, Seneca L Bessling, Webb Miller, Marcelo A Nobrega, Andrew S McCallion, and Ivan Ovcharenko. "Genome-wide identification of conserved regulatory function in diverged sequences." *Genome Research* 21.7 (July 2011), pp. 1139–1149. ISSN: 1549-5469. DOI: 10.1101/gr.119016.110.

[101] Cosmas D Arnold, Daniel Gerlach, Daniel Spies, Jessica A Matts, Yuliya A Sytnikova, Michaela Pagani, Nelson C Lau, and Alexander Stark. "Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution." *Nature Genetics* 46.7 (July 2014), pp. 685–692. DOI: 10.1038/ng.3009.

[102] D Yaffe, U Nudel, Y Mayer, and S Neuman. "Highly conserved sequences in the 3′ untranslated region of mRNAs coding for homologous proteins in distantly related species." *Nucleic Acids Research* 13.10 (May 1985), pp. 3723–3737. DOI: 10.1093/nar/13.10.3723.

[103] C Lemaire, R Heilig, and J L Mandel. "The chicken dystrophin cDNA: striking conservation of the C-terminal coding and 3′ untranslated regions between man and chicken." *The EMBO Journal* 7.13 (Dec. 1988), pp. 4157–4162.

[104] J P Rouault, C Samarut, L Duret, C Tessa, J Samarut, and J P Magaud. "Sequence analysis reveals that the BTG1 anti-proliferative gene is conserved throughout evolution in its coding and 3′ non-coding regions." *Gene* 129.2 (July 1993), pp. 303–306. DOI: 10.1016/0378-1119(93)90284-A.

[105] S Bagheri-Fam, C Ferraz, J Demaille, G Scherer, and D Pfeifer. "Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions." *Genomics* 78.1-2 (Nov. 2001), pp. 73–82. DOI: 10.1006/geno.2001.6648.

[106] Noël Ghanem, Olga Jarinova, Angel Amores, Qiaoming Long, Gary Hatch, Byung Keon Park, John L R Rubenstein, and Marc Ekker. "Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters." *Genome Research* 13.4 (Apr. 2003), pp. 533–543. DOI: 10.1101/gr.716103.

[107] Simona Santini, Jeffrey L Boore, and Axel Meyer. "Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters." *Genome Research* 13.6A (June 2003), pp. 1111–1122. DOI: 10.1101/gr.700503.

[108] Sol Katzman, Andrew D Kern, Gill Bejerano, Ginger Fewell, Lucinda Fulton, Richard K Wilson, Sofie R Salama, and David Haussler. "Human genome ultraconserved elements are ultraselected." *Science* 317.5840 (Aug. 2007), p. 915. ISSN: 1095-9203. DOI: 10.1126/science.1142430.

[109] Dimitris Polychronopoulos, Diamantis Sellis, and Yannis Almirantis. "Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics." *Plos One* 9.5 (May 2014), e95437. DOI: 10.1371/journal.pone.0095437.

[110] Hiroshi Kikuta et al. "Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates." *Genome Research* 17.5 (May 2007), pp. 545–555. ISSN: 1088-9051. DOI: 10.1101/gr.6086307.

[111] Altuna Akalin, David Fredman, Erik Arner, Xianjun Dong, Jan Christian Bryne, Harukazu Suzuki, Carsten O Daub, Yoshihide Hayashizaki, and Boris Lenhard. "Transcriptional features of genomic regulatory blocks." *Genome Biology* 10.4 (Apr. 2009), R38. DOI: 10.1186/gb-2009-10-4-r38.

[112] Nathan Harmston, Elizabeth Ing-Simmons, Ge Tan, Malcolm Perry, Matthias Merkenschlager, and Boris Lenhard. "Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation." *Nature Communications* 8.1 (Sept. 2017), p. 441. DOI: 10.1038/s41467-017-00524-5.

[113] Andrew C Nelson and Fiona C Wardle. "Conserved non-coding elements and cis regulation: actions speak louder than words." *Development* 140.7 (Apr. 2013), pp. 1385–1395. DOI: 10.1242/dev.084459.

[114] Dimitris Polychronopoulos, James W D King, Alexander J Nash, Ge Tan, and Boris Lenhard. "Conserved non-coding elements: developmental gene regulation meets genome organization." *Nucleic Acids Research* 45.22 (Dec. 2017), pp. 12611–12624. DOI: 10.1093/nar/gkx1074.

[115] Pehr Edman, Erik Högfeldt, Lars Gunnar Sillén, and Per-Olof Kinell. "Method for determination of the amino acid sequence in peptides." *Acta chemica Scandinavica* 4 (1950), pp. 283–293. ISSN: 0904-213X. DOI: 10.3891/acta.chem.scand.04-0283.

[116] R W Holley, J Apgar, G A Everett, J T Madison, M Marquisee, S H Merrill, J R Penswick, and A Zamir. "Structure of a ribonucleic acid." *Science* 147.3664 (Mar. 1965), pp. 1462–1465.

[117] Ray Wu and A.D. Kaiser. "Structure and base sequence in the cohesive ends of bacteriophage lambda DNA." *Journal of Molecular Biology* 35.3 (Jan. 1968), pp. 523–537. ISSN: 00222836. DOI: 10.1016/S0022-2836(68)80012-9.

[118] W Gilbert and A Maxam. "The nucleotide sequence of the lac operator." *Proceedings of the National Academy of Sciences of the United States of America* 70.12 (Dec. 1973), pp. 3581–3584. DOI: 10.1073/pnas.70.12.3581.

[119] A M Maxam and W Gilbert. "A new method for sequencing DNA." *Proceedings of the National Academy of Sciences of the United States of America* 74.2 (Feb. 1977), pp. 560–564. DOI: 10.1073/pnas.74.2.560.

[120] F Sanger, S Nicklen, and A R Coulson. "DNA sequencing with chain-terminating inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (Dec. 1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

[121] R Staden. "A strategy of DNA sequencing employing computer programs." *Nucleic Acids Research* 6.7 (June 1979), pp. 2601–2610. DOI: 10.1093/nar/6.7.2601.

[122] J Messing, R Crea, and P H Seeburg. "A system for shotgun DNA sequencing." *Nucleic Acids Research* 9.2 (Jan. 1981), pp. 309–321. DOI: 10.1093/nar/9.2.309.

[123] L M Smith, J Z Sanders, R J Kaiser, P Hughes, C Dodd, C R Connell, C Heiner, S B Kent, and L E Hood. "Fluorescence detection in automated DNA sequence analysis." *Nature* 321.6071 (June 1986), pp. 674–679. DOI: 10.1038/321674a0.

[124] David R Bentley et al. "Accurate whole human genome sequencing using reversible terminator chemistry." *Nature* 456.7218 (Nov. 2008), pp. 53–59. DOI: 10.1038/nature07517.

[125] John Eid et al. "Real-time DNA sequencing from single polymerase molecules." *Science* 323.5910 (Jan. 2009), pp. 133–138. ISSN: 1095-9203. DOI: 10.1126/science.1162986.

[126] Hagan Bayley. "Nanopore sequencing: from imagination to reality." *Clinical Chemistry* 61.1 (Jan. 2015), pp. 25–31. DOI: 10.1373/clinchem.2014.223016.

[127] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. "DNA sequencing at 40: past, present and future." *Nature* 550.7676 (Oct. 2017), pp. 345–353. DOI: 10.1038/nature24286.

[128] D R Hewish and L A Burgoyne. "Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease." *Biochemical and Biophysical Research Communications* 52.2 (May 1973), pp. 504–510. DOI: 10.1016/0006-291x(73)90740-7.

[129] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. "Chromatin accessibility and the regulatory epigenome." *Nature Reviews. Genetics* 20.4 (2019), pp. 207–220. ISSN: 1471-0056. DOI: 10.1038/s41576-018-0089-8.

[130] Hashem Koohy, Thomas A Down, and Tim J Hubbard. "Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme." *Plos One* 8.7 (July 2013), e69853. DOI: 10.1371/journal.pone.0069853.

[131] Peter J Sabo et al. "Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays." *Nature Methods* 3.7 (July 2006), pp. 511–518. DOI: 10.1038/nmeth890.

[132] Lingyun Song and Gregory E Crawford. "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells." *Cold Spring Harbor Protocols* 2010.2 (Feb. 2010), pdb.prot5384. DOI: 10.1101/pdb.prot5384.

[133] Sunil Gangadharan, Loris Mularoni, Jennifer Fain-Thornton, Sarah J Wheelan, and Nancy L Craig. "DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo." *Proceedings of the National Academy of Sciences of the United States of America* 107.51 (Dec. 2010), pp. 21966–21972. DOI: 10.1073/pnas.1016382107.

[134] M Ryan Corces et al. "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues." *Nature Methods* 14.10 (Oct. 2017), pp. 959–962. DOI: 10.1038/nmeth.4396.

[135] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. "Single-cell chromatin accessibility reveals principles of regulatory variation." *Nature* 523.7561 (July 2015), pp. 486–490. DOI: 10.1038/nature14590.

[136] Feng Yan, David R Powell, David J Curtis, and Nicholas C Wong. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis." *Genome Biology* 21.1 (Feb. 2020), p. 22. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1929-3.

[137] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. "High-resolution profiling of histone methylations in the human genome." *Cell* 129.4 (May 2007), pp. 823–837. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.05.009.

[138] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. "Genome-wide mapping of in vivo protein-DNA interactions." *Science* 316.5830 (June 2007), pp. 1497–1502. ISSN: 1095-9203. DOI: 10.1126/science.1141319.

[139] Clifford A Meyer and X Shirley Liu. "Identifying and mitigating bias in next-generation sequencing methods for chromatin biology." *Nature Reviews. Genetics* 15.11 (Nov. 2014), pp. 709–721. DOI: 10.1038/nrg3788.

[140] Laura Baranello, Fedor Kouzine, Suzanne Sanford, and David Levens. "ChIP bias as a function of cross-linking time." *Chromosome Research* 24.2 (2016), pp. 175–181. DOI: 10.1007/s10577-015-9509-1.

[141] Ho Sung Rhee and B Franklin Pugh. "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution." *Cell* 147.6 (Dec. 2011), pp. 1408–1419. ISSN: 1097-4172. DOI: 10.1016/j.cell.2011.11.013.

[142] Peter J Skene and Steven Henikoff. "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites." *eLife* 6 (Jan. 2017). DOI: 10.7554/{eLife}.21856.

[143] Terrence S Furey. "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions." *Nature Reviews. Genetics* 13.12 (Dec. 2012), pp. 840–852. DOI: 10.1038/nrg3306.

[144] A K Banerjee. "5'-terminal cap structure in eucaryotic messenger ribonucleic acids." *Microbiological reviews* 44.2 (June 1980), pp. 175–205.

[145] Hazuki Takahashi, Timo Lassmann, Mitsuyoshi Murata, and Piero Carninci. "5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing." *Nature Protocols* 7.3 (Feb. 2012), pp. 542–561. DOI: 10.1038/nprot.2012.005.

[146] Mitsuyoshi Murata, Hiromi Nishiyori-Sueki, Miki Kojima-Ishiyama, Piero Carninci, Yoshihide Hayashizaki, and Masayoshi Itoh. "Detecting expressed genes using CAGE." *Methods in Molecular Biology* 1164 (2014), pp. 67–85. DOI: 10.1007/978-1-4939-0805-9_7.

[147] Nevena Cvetesic, Harry G Leitch, Malgorzata Borkowska, Ferenc Müller, Piero Carninci, Petra Hajkova, and Boris Lenhard. "SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA." *Genome Research* 28.12 (Nov. 2018), pp. 1943–1956. ISSN: 1088-9051. DOI: 10.1101/gr.235937.118.

[148] Marek Gierliński et al. "Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment." *Bioinformatics* 31.22 (Nov. 2015), pp. 3625–3630. DOI: 10.1093/bioinformatics/btv425.

[149] O. Cerf. "A REVIEW tailing of survival curves of bacterial spores." *Journal of Applied Bacteriology* 42.1 (Feb. 1977), pp. 1–19. ISSN: 00218847. DOI: 10.1111/j.1365-2672.1977.tb00665.x.

[150] R W Makuch, D H Freeman, and M F Johnson. "Justification for the lognormal distribution as a model for blood pressure." *Journal of chronic diseases* 32.3 (1979), pp. 245–250. DOI: 10.1016/0021-9681(79)90070-5.

[151] Lawrence R. Rabiner and Biing Hwang Juang. "An introduction to hidden Markov models." *IEEE ASSP Magazine* 3 (1986), pp. 4–16.

[152] A Krogh, M Brown, I S Mian, K Sjölander, and D Haussler. "Hidden Markov models in computational biology. Applications to protein modeling." *Journal of Molecular Biology* 235.5 (Feb. 1994), pp. 1501–1531. DOI: 10.1006/jmbi.1994.1104.

[153] S R Eddy. "Hidden Markov models." *Current Opinion in Structural Biology* 6.3 (June 1996), pp. 361–365. DOI: 10.1016/s0959-440x(96)80056-x.

[154] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis*. Cambridge: Cambridge University Press, 1998. ISBN: 9780511790492. DOI: 10.1017/{CBO9780511790492}.

[155] Timo Koski. *Hidden markov models for bioinformatics*. Vol. 2. Computational Biology. Dordrecht: Springer Netherlands, 2001. ISBN: 978-1-4020-0136-9. DOI: 10.1007/978-94-010-0612-5.

[156] J. B. S. Haldane. "The fitting of binomial distributions." *Annals of Eugenics* 11.1 (Jan. 1941), pp. 179–181. ISSN: 20501420. DOI: 10.1111/j.1469-1809.1941.tb02283.x.

[157] Zoubin Ghahramani and Michael I. Jordan. "Factorial Hidden Markov Models." *Machine Learning* (1997).

[158] Wojciech Pieczynski. "Chaînes de markov triplet." *Comptes Rendus Mathematique* 335.3 (Jan. 2002), pp. 275–278. ISSN: 1631073X. DOI: 10.1016/S1631-{073X}(02)02462-7.

[159] Zoubin Ghahramani. "An Introduction to Hidden Markov Models and Bayesian Networks." *IJPRAI* 15 (Feb. 2001), pp. 9–42. DOI: 10.1142/S0218001401000836.

[160] MJ. Beal, Z. Ghahramani, and CE. Rasmussen. "The Infinite Hidden Markov Model." *Advances in Neural Information Processing Systems 14*. Max-Planck-Gesellschaft. Cambridge, MA, USA: MIT Press, Sept. 2002, pp. 577–584.

[161] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581. ISSN: 01621459.

[162] H Mamitsuka. "Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models." *Proteins* 33.4 (Dec. 1998), pp. 460–474. DOI: 10.1002/(sici)1097-0134(19981201)33:4\textless460::aid-prot2\textgreater3.0.co;2-m.

[163] Miklós Bóna. *A Walk Through Combinatorics: An Introduction to Enumeration and Graph Theory*. Second Edition. World Scientific, 2006. ISBN: 9812568859.

[164] Edsger W. Dijkstra. "A note on two problems in connexion with graphs." *Numerische Mathematik* 1.1 (Dec. 1959), pp. 269–271. ISSN: 0029-599X. DOI: 10.1007/{BF01386390}.

[165] Moshe Sniedovich. "Dijkstra's algorithm revisited: the dynamic programming connexion." eng. *Control and Cybernetics* 35.3 (2006), pp. 599–620.

[166] Feng Chen, Aaron J Mackey, Jeroen K Vermunt, and David S Roos. "Assessing performance of orthology detection strategies applied to eukaryotic genomes." *Plos One* 2.4 (Apr. 2007), e383. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000383.

[167] Noboru Jo Sakabe, Daniel Savic, and Marcelo A Nobrega. "Transcriptional enhancers in development and disease." *Genome Biology* 13.1 (Jan. 2012), p. 238. DOI: 10.1186/gb-2012-13-1-238.

[168] Mark Rebeiz and Miltos Tsiantis. "Enhancer evolution and the origins of morphological novelty." *Current Opinion in Genetics & Development* 45 (Aug. 2017), pp. 115–123. DOI: 10.1016/j.gde.2017.04.006.

[169] Michael Bulger and Mark Groudine. "Functional and mechanistic diversity of distal transcription enhancers." *Cell* 144.3 (Feb. 2011), pp. 327–339. DOI: 10.1016/j.cell.2011.01.024.

[170] Jennifer L Plank and Ann Dean. "Enhancer function: mechanistic and genome-wide insights come together." *Molecular Cell* 55.1 (July 2014), pp. 5–14. DOI: 10.1016/j.molcel.2014.06.015.

[171] Lily Li and Zeba Wunderlich. "An enhancer's length and composition are shaped by its regulatory task." *Frontiers in genetics* 8 (May 2017), p. 63. ISSN: 1664-8021. DOI: 10.3389/fgene.2017.00063.

[172] T. R. Gregory, J. A. Nicol, H. Tamm, B. Kullman, K. Kullman, I. J. Leitch, B. G. Murray, D. F. Kapraun, J. Greilhuber, and M. D. Bennett. "Eukaryotic genome size databases." *Nucleic Acids Res.* 35.Database issue (Jan. 2007), pp. D332–338.

[173] N Neznanov, A Umezawa, and R G Oshima. "A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice." *The Journal of Biological Chemistry* 272.44 (Oct. 1997), pp. 27549–27557. DOI: 10.1074/jbc.272.44.27549.

[174] Deborah I Ritter, Zhiqiang Dong, Su Guo, and Jeffrey H Chuang. "Transcriptional enhancers in protein-coding exons of vertebrate developmental genes." *Plos One* 7.5 (May 2012), e35202. DOI: 10.1371/journal.pone.0035202.

[175] Evgeny Z Kvon. "Using transgenic reporter assays to functionally characterize enhancers in animals." *Genomics* 106.3 (Sept. 2015), pp. 185–192. DOI: 10.1016/j.ygeno.2015.06.007.

[176] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A Pennacchio. "VISTA Enhancer Browser–a database of tissue-specific human enhancers." *Nucleic Acids Research* 35.Database issue (Jan. 2007), pp. D88–92. DOI: 10.1093/nar/gkl822.

[177] Jamie C Kwasnieski, Ilaria Mogno, Connie A Myers, Joseph C Corbo, and Barak A Cohen. "Complex effects of nucleotide variants in a mammalian cis-regulatory element." *Proceedings of the National Academy of Sciences of the United States of America* 109.47 (Nov. 2012), pp. 19498–19503. DOI: 10.1073/pnas.1210678109.

[178] Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, and Alexander Stark. "Genome-wide quantitative enhancer activity maps identified by STARR-seq." *Science* 339.6123 (Mar. 2013), pp. 1074–1077. DOI: 10.1126/science.1232542.

[179] Waseem Akhtar, Johann de Jong, Alexey V Pindyurin, Ludo Pagie, Wouter Meuleman, Jeroen de Ridder, Anton Berns, Lodewyk F A Wessels, Maarten van Lohuizen, and Bas van Steensel. "Chromatin position effects assayed by thousands of reporters integrated in parallel." *Cell* 154.4 (Aug. 2013), pp. 914–927. DOI: 10.1016/j.cell.2013.07.018.

[180] Matthew Murtha et al. "FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells." *Nature Methods* 11.5 (May 2014), pp. 559–565. DOI: 10.1038/nmeth.2885.

[181] Diane E Dickel et al. "Function-based identification of mammalian enhancers using site-specific integration." *Nature Methods* 11.5 (May 2014), pp. 566–571. DOI: 10.1038/nmeth.2886.

[182] Fumitaka Inoue and Nadav Ahituv. "Decoding enhancers using massively parallel reporter assays." *Genomics* 106.3 (Sept. 2015), pp. 159–164. DOI: 10.1016/j.ygeno.2015.06.005.

[183] Christopher Fletez-Brant, Dongwon Lee, Andrew S McCallion, and Michael A Beer. "kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets." *Nucleic Acids Research* 41.Web Server issue (July 2013), W544–56. DOI: 10.1093/nar/gkt519.

[184] Bin Liu, Longyun Fang, Ren Long, Xun Lan, and Kuo-Chen Chou. "iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition." *Bioinformatics* 32.3 (Feb. 2016), pp. 362–369. DOI: 10.1093/bioinformatics/btv604.

[185] Bite Yang, Feng Liu, Chao Ren, Zhangyi Ouyang, Ziwei Xie, Xiaochen Bo, and Wenjie Shu. "BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone." *Bioinformatics* 33.13 (July 2017), pp. 1930–1936. DOI: 10.1093/bioinformatics/btx105.

[186] Anand Pratap Singh, Sarthak Mishra, and Suraiya Jabin. "Sequence based prediction of enhancer regions from DNA random walk." *Scientific Reports* 8.1 (Oct. 2018), p. 15912. DOI: 10.1038/s41598-018-33413-y.

[187] Nicole Rusk. "Predicting enhancers by their sequence." *Nature Methods* 11.6 (June 2014), pp. 606–607. ISSN: 1548-7091. DOI: 10.1038/nmeth.2987.

[188] S U Kass, N Landsberger, and A P Wolffe. "DNA methylation directs a time-dependent repression of transcription initiation." *Current Biology* 7.3 (Mar. 1997), pp. 157–165. DOI: 10.1016/s0960-9822(97)70086-1.

[189] Florence K Crary-Dooley, Mitchell E Tam, Keith W Dunaway, Irva Hertz-Picciotto, Rebecca J Schmidt, and Janine M LaSalle. "A comparison of existing global DNA methylation assays to low-coverage whole-genome bisulfite sequencing for epidemiological studies." *Epigenetics* 12.3 (Mar. 2017), pp. 206–214. DOI: 10.1080/15592294.2016.1276680.

[190] Masako Suzuki, Will Liao, Frank Wos, Andrew D Johnston, Justin DeGrazia, Jennifer Ishii, Toby Bloom, Michael C Zody, Soren Germer, and John M Greally. "Whole-genome bisulfite sequencing with improved accuracy and cost." *Genome Research* 28.9 (Aug. 2018), pp. 1364–1371. DOI: 10.1101/gr.232587.117.

[191] Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis." *Nucleic Acids Research* 33.18 (Oct. 2005), pp. 5868–5877. DOI: 10.1093/nar/gki901.

[192] Ulf Andersson Ørom et al. "Long noncoding RNAs with enhancer-like function in human cells." *Cell* 143.1 (Oct. 2010), pp. 46–58. ISSN: 1097-4172. DOI: 10.1016/j.cell.2010.09.001.

[193] Marcelo A Nobrega, Ivan Ovcharenko, Veena Afzal, and Edward M Rubin. "Scanning human gene deserts for long-range enhancers." *Science* 302.5644 (Oct. 2003), p. 413. ISSN: 1095-9203. DOI: 10.1126/science.1088328.

[194] Len A Pennacchio et al. "In vivo enhancer analysis of human conserved non-coding sequences." *Nature* 444.7118 (Nov. 2006), pp. 499–502. ISSN: 1476-4687. DOI: 10.1038/nature05295.

[195] Kerstin Lindblad-Toh et al. "A high-resolution map of human evolutionary constraint using 29 mammals." *Nature* 478.7370 (Oct. 2011), pp. 476–482. ISSN: 1476-4687. DOI: 10.1038/nature10530.

[196] Dimitrios Kleftogiannis, Panos Kalnis, and Vladimir B Bajic. "Progress and challenges in bioinformatics approaches for enhancer identification." *Briefings in Bioinformatics* 17.6 (2016), pp. 967–979. DOI: 10.1093/bib/bbv101.

[197] Leonard Whye Kit Lim, Hung Hui Chung, Yee Ling Chong, and Nung Kion Lee. "A survey of recently emerged genome-wide computational enhancer predictor tools." *Computational biology and chemistry* 74 (June 2018), pp. 132–141. DOI: 10.1016/j.compbiolchem.2018.03.019.

[198] Manolis Kellis et al. "Defining functional DNA elements in the human genome." *Proceedings of the National Academy of Sciences of the United States of America* 111.17 (Apr. 2014), pp. 6131–6138. DOI: 10.1073/pnas.1318948111.

[199] Alessandro Mammana and Johannes Helmuth. "bamsignals: Extract read count signals from bam files." R package version 1.12.1. 2016.

[200] Benedikt Zacher, Margaux Michel, Björn Schwalb, Patrick Cramer, Achim Tresch, and Julien Gagneur. "Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by genostan." *Plos One* 12.1 (Jan. 2017), e0169249. DOI: 10.1371/journal.pone.0169249.

[201] Kyoung-Jae Won, Xian Zhang, Tao Wang, Bo Ding, Debasish Raha, Michael Snyder, Bing Ren, and Wei Wang. "Comparative annotation of functional regions in the human genome using epigenomic data." *Nucleic Acids Research* 41.8 (Apr. 2013), pp. 4423–4432. DOI: 10.1093/nar/gkt143.

[202] Gary Hon, Bing Ren, and Wei Wang. "ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome." *PLoS Computational Biology* 4.10 (Oct. 2008), e1000201. DOI: 10.1371/journal.pcbi.1000201.

[203] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. "Unsupervised pattern discovery in human chromatin structure through genomic segmentation." *Nature Methods* 9.5 (Mar. 2012), pp. 473–476. DOI: 10.1038/nmeth.1937.

[204] Nisha Rajagopal, Wei Xie, Yan Li, Uli Wagner, Wei Wang, John Stamatoyannopoulos, Jason Ernst, Manolis Kellis, and Bing Ren. "RFECS: a random-forest based algorithm for enhancer identification from chromatin state." *PLoS Computational Biology* 9.3 (Mar. 2013), e1002968. DOI: 10.1371/journal.pcbi.1002968.

[205] Feng Liu, Hao Li, Chao Ren, Xiaochen Bo, and Wenjie Shu. "PEDLA: predicting enhancers with a deep learning-based algorithmic framework." *Scientific Reports* 6 (June 2016), p. 28517. DOI: 10.1038/srep28517.

[206] Lan T M Dao et al. "Genome-wide characterization of mammalian promoters with distal enhancer functions." *Nature Genetics* 49.7 (July 2017), pp. 1073–1081. DOI: 10.1038/ng.3884.

[207] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)." *Genome Biology* 9.9 (Sept. 2008), R137. DOI: 10.1186/gb-2008-9-9-r137.

[208] Fan Hsu, W James Kent, Hiram Clawson, Robert M Kuhn, Mark Diekhans, and David Haussler. "The UCSC known genes." *Bioinformatics* 22.9 (May 2006), pp. 1036–1046. DOI: 10.1093/bioinformatics/btl048.

[209] Tin Y Lam and Irmtraud M Meyer. "Efficient algorithms for training the parameters of hidden Markov models using stochastic expectation maximization (EM) training and Viterbi training." *Algorithms for Molecular Biology* 5 (Dec. 2010), p. 38. DOI: 10.1186/1748-7188-5-38.

[210] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489.7414 (Sept. 2012), pp. 57–74. DOI: 10.1038/nature11247.

[211] Alessandro Mammana and Ho-Ryun Chung. "Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome." *Genome Biology* 16 (July 2015), p. 151. DOI: 10.1186/s13059-015-0708-z.

[212] Tianshun Gao, Bing He, Sheng Liu, Heng Zhu, Kai Tan, and Jiang Qian. "EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types." *Bioinformatics* 32.23 (Dec. 2016), pp. 3543–3551. DOI: 10.1093/bioinformatics/btw495.

[213] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017.

[214] D Eddelbuettel and R François. "rcpp : seamless R and C++ Integration." *Journal of statistical software* 40.8 (2011). ISSN: 1548-7660. DOI: 10.18637/jss.v040.i08.

[215] B M Bolstad, R A Irizarry, M Astrand, and T P Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics* 19.2 (Jan. 2003), pp. 185–193. DOI: 10.1093/bioinformatics/19.2.185.

[216] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data sets–update." *Nucleic Acids Research* 41.Database issue (Jan. 2013), pp. D991–5. DOI: 10.1093/nar/gks1193.

[217] Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. "CrossMap: a versatile tool for coordinate conversion between genome assemblies." *Bioinformatics* 30.7 (Apr. 2014), pp. 1006–1007. DOI: 10.1093/bioinformatics/btt730.

[218] Adam Siepel et al. "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." *Genome Research* 15.8 (Aug. 2005), pp. 1034–1050. ISSN: 1088-9051. DOI: 10.1101/gr.3715005.

[219] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. "The sequence read archive." *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D19–21. DOI: 10.1093/nar/gkq1019.

[220] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* 25.14 (July 2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[221] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

[222] H M Chan and N B La Thangue. "p300/CBP proteins: HATs for transcriptional bridges and scaffolds." *Journal of Cell Science* 114.Pt 13 (July 2001), pp. 2363–2373.

[223] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization." *Nature Methods* 9.3 (Mar. 2012), pp. 215–216. DOI: 10.1038/nmeth.1906.

[224] Yupeng He et al. "Improved regulatory element prediction based on tissue-specific local epigenomic signatures." *Proceedings of the National Academy of Sciences of the United States of America* 114.9 (Feb. 2017), E1633–E1640. DOI: 10.1073/pnas.1618353114.

[225] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." *Molecular Cell* 38.4 (May 2010), pp. 576–589. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2010.05.004.

[226] Bum-Kyu Lee and Vishwanath R Iyer. "Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation." *The Journal of Biological Chemistry* 287.37 (Sept. 2012), pp. 30906–30913. DOI: 10.1074/jbc.R111.324962.

[227] Ya Guo et al. "CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function." *Cell* 162.4 (Aug. 2015), pp. 900–910. ISSN: 00928674. DOI: 10.1016/j.cell.2015.07.038.

[228] J. Huang, K. Li, W. Cai, X. Liu, Y. Zhang, S. H. Orkin, J. Xu, and G. C. Yuan. "Dissecting super-enhancer hierarchy based on chromatin interactions." *Nat Commun* 9.1 (Mar. 2018), p. 943.

[229] A. Sharifi-Zarchi et al. "DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism." *BMC Genomics* 18.1 (Dec. 2017), p. 964.

[230] A. Meissner et al. "Genome-scale DNA methylation maps of pluripotent and differentiated cells." *Nature* 454.7205 (Aug. 2008), pp. 766–770.

[231] Shyam Prabhakar, Francis Poulin, Malak Shoukry, Veena Afzal, Edward M Rubin, Olivier Couronne, and Len A Pennacchio. "Close sequence comparisons are sufficient to identify human cis-regulatory elements." *Genome Research* 16.7 (July 2006), pp. 855–863. ISSN: 1088-9051. DOI: 10.1101/gr.4717506.

[232] Dominic Schmidt et al. "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding." *Science* 328.5981 (May 2010), pp. 1036–1040. ISSN: 1095-9203. DOI: 10.1126/science.1186176.

[233] Thomas S Becker and Boris Lenhard. "The random versus fragile breakage models of chromosome evolution: a matter of resolution." *Molecular Genetics and Genomics* 278.5 (Nov. 2007), pp. 487–491. DOI: 10.1007/s00438-007-0287-0.

[234] Sudhir Kumar, Glen Stecher, Michael Suleski, and S Blair Hedges. "Timetree: A resource for timelines, timetrees, and divergence times." *Molecular Biology and Evolution* 34.7 (July 2017), pp. 1812–1819. DOI: 10.1093/molbev/msx116.

[235] Byrappa Venkatesh et al. "Elephant shark genome provides unique insights into gnathostome evolution." *Nature* 505.7482 (Jan. 2014), pp. 174–179. DOI: 10.1038/nature12826.

[236] Nathan Harmston, Anja Baresic, and Boris Lenhard. "The mystery of extreme non-coding conservation." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368.1632 (Dec. 2013), p. 20130021. DOI: 10.1098/rstb.2013.0021.

[237] Carlos Gómez-Marín et al. "Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders." *Proceedings of the National Academy of Sciences of the United States of America* 112.24 (June 2015), pp. 7542–7547. DOI: 10.1073/pnas.1505463112.

[238] Jan Krefting, Miguel A Andrade-Navarro, and Jonas Ibn-Salem. "Evolutionary stability of topologically associating domains is associated with conserved gene regulation." *BMC Biology* 16.1 (Aug. 2018), p. 87. ISSN: 1741-7007. DOI: 10.1186/s12915-018-0556-x.

[239] Aimée Zuniga. "Next generation limb development and evolution: old questions, new perspectives." *Development* 142.22 (Nov. 2015), pp. 3810–3820. DOI: 10.1242/dev.125757.

[240] Jennifer A. Clack. "The Fin to Limb Transition: New Data, Interpretations, and Hypotheses from Paleontology and Developmental Biology." *Annual review of earth and planetary sciences* 37.1 (May 2009), pp. 163–179. ISSN: 0084-6597. DOI: 10.1146/annurev.earth.36.031207.124146.

[241] Thomas W P Wood and Tetsuya Nakamura. "Problems in Fish-to-Tetrapod Transition: Genetic Expeditions Into Old Specimens." *Frontiers in cell and developmental biology* 6 (July 2018), p. 70. DOI: 10.3389/fcell.2018.00070.

[242] Haihan Tan, Daria Onichtchouk, and Cecilia Winata. "DANIO-CODE: Toward an Encyclopedia of DNA Elements in Zebrafish." *Zebrafish* 13.1 (Feb. 2016), pp. 54–60. DOI: 10.1089/zeb.2015.1179.

[243] Matthias Hörtenhuber, Abdul K Mukarram, Marcus H Stoiber, James B Brown, and Carsten O Daub. "*-DCC: A platform to collect, annotate, and explore a large variety of sequencing experiments." *GigaScience* 9.3 (Mar. 2020). DOI: 10.1093/gigascience/giaa024.

[244] Ozren Bogdanović et al. "Active DNA demethylation at enhancers during the vertebrate phylotypic period." *Nature Genetics* 48.4 (Apr. 2016), pp. 417–426. DOI: 10.1038/ng.3522.

[245] Elisa de la Calle Mustienes, Jose Luis Gómez-Skarmeta, and Ozren Bogdanović. "Genome-wide epigenetic cross-talk between DNA methylation and H3K27me3 in zebrafish embryos." *Genomics data* 6 (Dec. 2015), pp. 7–9. DOI: 10.1016/j.gdata.2015.07.020.

[246] Christina Paliou et al. "Preformed chromatin topology assists transcriptional robustness of Shh during limb development." *Proceedings of the National Academy of Sciences of the United States of America* 116.25 (June 2019), pp. 12390–12399. DOI: 10.1073/pnas.1900672116.

[247] Guillaume Andrey et al. "Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding." *Genome Research* 27.2 (2017), pp. 223–233. DOI: 10.1101/gr.213066.116.

[248] Yupeng He et al. "Spatiotemporal DNA methylome dynamics of the developing mouse fetus." *Nature* 583.7818 (July 2020), pp. 752–759. ISSN: 0028-0836. DOI: 10.1038/s41586-020-2119-x.

[249] V Hamburger and H L Hamilton. "A series of normal stages in the development of the chick embryo." *Journal of Morphology* 88.1 (Jan. 1951), pp. 49–92. DOI: 10.1002/jmor.1050880104.

[250] V Hamburger and H L Hamilton. "A series of normal stages in the development of the chick embryo. 1951." *Developmental Dynamics* 195.4 (Dec. 1992), pp. 231–272. DOI: 10.1002/aja.1001950404.

[251] Alexandra Despang et al. "Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture." *Nature Genetics* 51.8 (July 2019), pp. 1263–1271. ISSN: 1061-4036. DOI: 10.1038/s41588-019-0466-z.

[252] Ge Tan, Dimitris Polychronopoulos, and Boris Lenhard. "CNEr: A toolkit for exploring extreme noncoding conservation." *PLoS Computational Biology* 15.8 (Aug. 2019), e1006940. DOI: 10.1371/journal.pcbi.1006940.

[253] Stella M K Glasauer and Stephan C F Neuhauss. "Whole-genome duplication in teleost fishes and its evolutionary consequences." *Molecular Genetics and Genomics* 289.6 (Dec. 2014), pp. 1045–1060. DOI: 10.1007/s00438-014-0889-2.

[254] Martín D Ezcurra, Torsten M Scheyer, and Richard J Butler. "The origin and early evolution of Sauria: reassessing the permian Saurian fossil record and the timing of the crocodile-lizard divergence." *Plos One* 9.2 (Feb. 2014), e89165. DOI: 10.1371/journal.pone.0089165.

[255] Deborah I Ritter, Qiang Li, Dennis Kostka, Katherine S Pollard, Su Guo, and Jeffrey H Chuang. "The importance of being cis: evolution of orthologous fish and mammalian enhancer activity." *Molecular Biology and Evolution* 27.10 (Oct. 2010), pp. 2322–2332. DOI: 10.1093/molbev/msq128.

[256] Adrian P Bracken, Nikolaj Dietrich, Diego Pasini, Klaus H Hansen, and Kristian Helin. "Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions." *Genes & Development* 20.9 (May 2006), pp. 1123–1136. DOI: 10.1101/gad.381706.

[257] Laurie A Boyer et al. "Polycomb complexes repress developmental regulators in murine embryonic stem cells." *Nature* 441.7091 (May 2006), pp. 349–353. ISSN: 1476-4687. DOI: 10.1038/nature04733.

[258] Feng Tie, Rakhee Banerjee, Alina R Saiakhova, Benny Howard, Kelsey E Monteith, Peter C Scacheri, Michael S Cosgrove, and Peter J Harte. "Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing." *Development* 141.5 (Mar. 2014), pp. 1129–1139. DOI: 10.1242/dev.102392.

[259] Andrei Kuzmichev, Kenichi Nishioka, Hediye Erdjument-Bromage, Paul Tempst, and Danny Reinberg. "Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein." *Genes & Development* 16.22 (Nov. 2002), pp. 2893–2905. DOI: 10.1101/gad.1035902.

[260] Karin J Ferrari, Andrea Scelfo, Sriganesh Jammula, Alessandro Cuomo, Iros Barozzi, Alexandra Stützer, Wolfgang Fischle, Tiziana Bonaldi, and Diego Pasini. "Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity." *Molecular Cell* 53.1 (Jan. 2014), pp. 49–62. DOI: 10.1016/j.molcel.2013.10.030.

[261] Daniel J Grau, Brad A Chapman, Joe D Garlick, Mark Borowsky, Nicole J Francis, and Robert E Kingston. "Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge." *Genes & Development* 25.20 (Oct. 2011), pp. 2210–2221. DOI: 10.1101/gad.17288211.

[262] Bernd Schuettengruber, Henri-Marc Bourbon, Luciano Di Croce, and Giacomo Cavalli. "Genome regulation by polycomb and trithorax: 70 years and counting." *Cell* 171.1 (Sept. 2017), pp. 34–57. ISSN: 00928674. DOI: 10.1016/j.cell.2017.08.002.

[263] Rory T Coleman and Gary Struhl. "Causal role for inheritance of H3K27me3 in maintaining the OFF state of a Drosophila HOX gene." *Science* 356.6333 (Apr. 2017). DOI: 10.1126/science.aai8236.

[264] Koustav Pal, Mattia Forcato, and Francesco Ferrari. "Hi-C analysis: from data generation to integration." *Biophysical reviews* 11.1 (Feb. 2019), pp. 67–78. DOI: 10.1007/s12551-018-0489-1.

[265] Laura Vian et al. "The energetics and physiological impact of cohesin extrusion." *Cell* 173.5 (May 2018), 1165–1178.e20. ISSN: 00928674. DOI: 10.1016/j.cell.2018.03.072.

[266] Katerina Kraft et al. "Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations." *Nature Cell Biology* 21.3 (Feb. 2019), pp. 305–310. ISSN: 1465-7392. DOI: 10.1038/s41556-019-0273-x.

[267] Vanja Haberle. *seqPattern: Visualising oligonucleotide patterns and motif occurrences across a set of sorted sequences*. R package version 1.16.0. 2019.

[268] Oriol Fornes et al. "JASPAR 2020: update of the open-access database of transcription factor binding profiles." *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D87–D92. ISSN: 0305-1048. DOI: 10.1093/nar/gkz1001.

[269] Edgar Wingender. "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation." *Briefings in Bioinformatics* 9.4 (July 2008), pp. 326–332. DOI: 10.1093/bib/bbn016.

[270] Brandon Milholland, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. "Differences between germline and somatic mutation rates in humans and mice." *Nature Communications* 8 (May 2017), p. 15183. DOI: 10.1038/ncomms15183.

[271] Anna Ramisch et al. "CRUP: a comprehensive framework to predict condition-specific regulatory units." *Genome Biology* 20.1 (Nov. 2019), p. 227. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1860-7.

[272] Xi Chen et al. "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." *Cell* 133.6 (June 2008), pp. 1106–1117. ISSN: 1097-4172. DOI: 10.1016/j.cell.2008.04.043.

[273] Yin Shen et al. "A map of the cis-regulatory sequences in the mouse genome." *Nature* 488.7409 (Aug. 2012), pp. 116–120. DOI: 10.1038/nature11243.

[274] Pengfei Yu, Shu Xiao, Xiaoyun Xin, Chun-Xiao Song, Wei Huang, Darina McDee, Tetsuya Tanaka, Ting Wang, Chuan He, and Sheng Zhong. "Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation." *Genome Research* 23.2 (Feb. 2013), pp. 352–364. DOI: 10.1101/gr.144949.112.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

A	adenine
ATAC-seq	assay for transposase-accessible chromatin using sequencing
AUPRC	area under the precision recall curve
bp	base pair
BWA	Burrows-Wheeler Alignment
C	cytosine
CAGE	cap analysis gene expression
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CNE	conserved non-coding element
CTCF	CCCTC-binding factor
DA	Dijkstra's Shortest Path Algorithm
DBD	DNA-binding domain
DHS	DNase I hypersensitive site
DMR	differentially methylated region

DNA	desoxyribonucleic acid
DNase I	deoxyribonuclease I
DNase-seq	DNase I hypersensitive site sequencing
E step	Expectation step
eHMM	enhancer Hidden Markov Model
EM	expectation-maximization
ENCODE	Encyclopedia of DNA Elements
EpiCSeg	Epigenome Count-based Segmentation
eRNA	enhancer **RNA**
ESC	embryonic stem cells
FANTOM	Functional Annotation of the Mammalian Genome
G	guanine
Gbp	giga base pairs
GEO	Gene Expression Omnibus
GRB	genomic regulatory block
GTF	general transcription factor
H3K27	lysine 27 at histone 3
H3K27ac	acetylation of histone 3, lysine 27
H3K27me3	trimethylation of histone 3, lysine 4
H3K36me3	trimethylation of histone 3, lysine 36
H3K4	lysine 4 at histone 3
H3K4me1	monomethylation of histone 3, lysine 4
H3K4me1/3	mono- and trimethylation of histone 3, lysine 4
H3K4me2	dimethylation of histone 3, lysine 4
H3K4me3	trimethylation of histone 3, lysine 4
HAT	histone acetyltransferase
HH25	Hamburger-Hamilton stage 25
HM	histone modification
HMM	hidden Markov model
HTS	high-throughput sequencing

IPP	Independent Point Projection
kbp	kilo base pairs
KS	Kolmogorov-Smirnov
LAD	lamina associated domain
LMR	low-methylated region
lncRNA	long non-coding RNA
M step	Maximization step
MAP	maximum a posteriori estimation
Mbp	mega base pairs
MDa	megadalton
MeDIP-seq	Methylated DNA immunoprecipitation followed by sequencing
miRNA	microRNA
MLE	maximum likelihood estimation
MPRA	massively parallel reporter assay
mRNA	messenger RNA
mya	million years ago
myr	million years
NCBI	National Center for Biotechnology Information
NGS	next generation sequencing
PcG	Polycomb
PCR	polymerase chain reaction
PRC	Polycomb repressive complex
RNA	ribonucleic acid
RNAP II	RNA Polymerase II
ROC	receiver operating characteristic
RRBS	reduced representation bisulfite sequencing

rRNA	ribosomal RNA
SAPP	Syntenic Anchor Point Propagation
scATAC-seq	single-cell ATAC-seq
siRNA	small interfering RNA
SMRT	single-molecule real-time sequencing
SRA	Sequence Read Archive
T	thymine
TAD	topologically associated domain
TF	transcription factor
TFBS	transcription factor binding site
tRNA	transfer RNA
TrxG	Trithorax
TSS	transcription start site
U	uracil
UCSC	University of California Santa Cruz
UTR	untranslated region
WGBS	whole-genome bisulfite sequencing
WGD	whole genome duplication

CURRICULUM VITAE

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

SUMMARY

In the last several decades, the field of molecular biology has made substantial progress in deciphering the many facets of gene expression and regulation. The parallel development of experimental methods, especially high-throughput, greatly contributed to new possibilities for studying functional elements of the genome such as promoters and enhancers. In this thesis, I investigate the characterization and evolutionary conservation of enhancers.

First, I present **eHMM**, a method that uses a supervised **HMM** with a constrained underlying Markov chain that incorporates prior biological knowledge about the molecular structure of enhancers in a dynamic model to predict heterogeneous enhancers of variable sizes on the basis of a minimal set of features. I demonstrate **eHMM**'s prediction performance using different validation setups within and across data sets, tissues and developmental stages and analyze genome-wide predictions in terms of functional genomic and epigenomic features, spatial accuracy and susceptibility for false-positive results.

Second, I investigate functional evolutionary conservation of enhancers in absence of detectable sequence conservation. For that, I introduce the concept of using multiple sets of pairwise alignments that allow moving through a species graph in order to produce accurate projections of non-alignable genomic regions between two species with large evolutionary distances. To that end, I present the methods **IPP** and **SAPP** that approach the task under slightly different aspects.

IPP projects individual genomic point coordinates from one species onto another by interpolating their position between two alignable sequences, so-called anchor points. Instead of using only direct alignments between the two species in question, **IPP** implements the choice of an optimal set of bridging species that maximizes projection accuracy. I demonstrate **IPP**'s projection accuracy compared to using direct alignments, propose functional conservation to be a universal phenomenon and identify individual occurrences of functional orthologs beyond sequence conservation.

SAPP propagates anchor points rather than projecting genomic points in a fashion that minimizes resulting anchor spans in the target species. By that, it respects the conservation of synteny and provides maximally narrowed search spaces for analyzing enhancer equivalence between two species.

Together, the work presented in this thesis aims at adding to our current understanding about the identity and the evolutionary properties of enhancers.

ZUSAMMENFASSUNG

Die Molekularbiologie hat in den letzten Jahrzehnten grosse Fortschritte im Bereich der Genexpression und deren Regulation gemacht. Die parallele Entwicklung von experimentellen Methoden, speziell 'high-throughput', hat neue Möglichkeiten geschaffen um funktionelle Elemente des Genoms wie Promotoren und Enhancer zu studieren. In dieser Dissertation untersuche ich die Charakterisierung sowie die evolutionäre Konservierung von Enhancern.

Zunächst präsentiere ich **eHMM**, eine Methode zur Bestimmung von heterogenen Enhancern mit variablen Grössen anhand minimaler Merkmale mittels überwachtem **HMM** mit eingeschränkter Markov-Kette, die biologische Kenntnisse über die molekulare Struktur von Enhancern in einem dynamischen Modell integriert. Ich validiere **eHMM**s Enhancerklassifizierung sowohl innerhalb als auch zwischen Datensätzen, Geweben und Entwicklungsstadien. Ausserdem analysiere ich genomweite Enhancervorhersagen betreffend funktioneller genomischen und epigenomischen Merkmalen, räumlicher Präzision und der Anfälligkeit für falsch-positive Resultate.

Des Weiteren untersuche ich die funktionelle evolutionäre Konservierung von Enhancern in Fällen fehlender Sequenzkonservierung. Ich stelle dafür ein Konzept vor, das exakte Projektionen nicht-alinierbarer Regionen zwischen Spezies mit grossen evolutionären Distanzen ermöglicht und präsentiere zu diesem Zweck die Methoden **IPP** und **SAPP**, die diese Aufgabe unter leicht verschiedenen Aspekten angehen.

IPP projiziert individuelle genomische Punktkoordinaten zwischen zwei Spezies mittels Interpolation derer Positionen zwischen zwei alinierbaren Ankerpunkten. **IPP** implementiert die optimierte Auswahl sogenannter "Brückenspezies" zur Maximierung der Projektionsgenauigkeit, welche ich im Vergleich mit der Verwendung von lediglich direkten Alignments zwischen den betreffenden Spezies analysiere. Darüber hinaus schlage ich funktionelle Konservierung als universelles Phänomen vor und identifiziere individuelle funktionelle Orthologe jenseits der Konservierung der Sequenz.

Im Gegensatz dazu propagiert **SAPP** die Ankerpunkte, so dass die Abstände der resultierenden Ankerpunkten in der Zielspezies minimiert werden. Dadurch wird die Konservierung von Syntenie berücksichtigt und die Analyse von Enhanceräquivalenz in maximal reduzierten Suchfeldern ermöglicht.

Insgesamt leistet die in dieser Dissertation präsentierte Arbeit einen Beitrag zum gegenwärtigen Verständnis der Identität und den evolutionären Eigenschaften von Enhancern.

SELBSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, Tobias Zehnder, gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht. Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Berlin, September 2020

Tobias Zehnder