ENHANCED ALGORITHMS TO DETECT AND CHARACTERIZE CONSERVED REGULATORY SEQUENCES

by

Jin Woo Oh

A dissertation submitted to Johns Hopkins University in conformity with the requirements for

the degree of Doctor of Philosophy

Baltimore, Maryland

April 2024

Abstract

Mutations in gene regulatory elements are associated with increased risk for many complex genetic diseases such as schizophrenia, diabetes, and cancer. However, the degenerate sequence structures of regulatory elements and their complex contribution to gene expression pose great challenges to understanding pathogenesis induced by regulatory mutations. To decipher the regulatory genome, advancement in both experimental and computational methods to characterize gene regulatory elements is indispensable. This dissertation details my contribution to this effort, with emphasis on the computational aspect.

First, I collaborated with members of the ENCODE and IGVF consortia to functionally characterize diverse gene regulatory elements through systematic epigenetic perturbation with CRISPRi. For example, we screened putative enhancers of core transcription factors that drive differentiation of embryonic stem cells to definitive endoderm (e.g., SOX17, GATA6), and discovered that enhancers modulate the speed of cell differentiation through enhancer strength and redundancy. Further, we quantitatively analyzed diverse CRISPR screen methodologies, and identified important principles of non-coding CRISPR screens that will streamline future efforts for enhancer functional characterization.

Second, to facilitate functional characterization of regulatory elements through model species, we developed a novel genome alignment algorithm designed to identify conserved distal enhancers. Biological functions and variant impacts of human enhancers are often tested using their conserved orthologous counterparts in model species such as mice. However, due to both rapid evolution of enhancers and computational limitations, mapping conserved distal enhancers between distant

mammals remains a challenge. To improve upon existing computational methods, we developed a novel genome alignment algorithm using gapped-kmer sequence features, which have previously been shown to effectively model regulatory sequences. We comprehensively evaluated the novel algorithm, *gkm-align*, using thousands of DNase-seq data generated from diverse human and mouse cell/tissues. Through this expansive dataset, we observed intriguingly high level of variation in enhancer conservation across distinct cell/tissues, which is explainable through association with transposable elements. This observation provides a quantitative support to the notion that enhancer evolution and transcriptional rewiring may be driven by transposable elements that carry human transcription factor binding sequences. We show that while cell-specific regulatory vocabulary is conserved, enhancers evolve more rapidly than other genomic elements such as promoters and CTCF binding sites. In spite of the rapid evolution, *gkm-align* discovers more than 20,000 novel enhancers with conserved epigenetic profiles, missed by standard alignment methods.

Thesis readers

Dr. Michael A. Beer (Primary Advisor)
Professor
Department of Biomedical Engineering and Genetic Medicine
Johns Hopkins University

Dr. Michael C. Schatz
Professor
Department of Computer Science and Biology
Johns Hopkins University

Dr. Patrick Cahan
Associate Professor
Department of Biomedical Engineering and Molecular Biology & Genetics
Johns Hopkins University

Dedication

This thesis is dedicated to my family for their love and support.

Acknowledgements

First, I would like to express my gratitude to my advisor and mentor, Dr. Michael Beer, for his unwavering support and guidance throughout my scientific journey. I first met him during my undergraduate years at Johns Hopkins through courses he taught. The genuine passion for science he exhibited in each lecture deeply resonated with me, introducing me to the beauty of seeking simple models and explanations for complex phenomena teeming in the natural world. After each lecture, he was always eager to discuss with me intricate details of the experiments and theories he introduced, and his enthusiasm for scientific discussion persisted throughout my PhD training in his lab. I am also grateful to him for providing with me exciting scientific opportunities that made this dissertation possible. I thank all the members of the Beer lab, including Dustin Shigaki, Milad Razavi-Mohseni, Gary Yang, Wang Xi, and Andrew Rojnuckarin, for being supportive team members and good friends.

I would also like to thank my thesis committee members, Dr. Michael Schatz and Dr. Cahan, for supporting my scientific training. I first met them in their courses in comparative genomics and computational stem cell biology, both of which are some of the core topics I will discuss in the dissertation. I would like to thank them for introducing me to these interesting fields of biological research and for providing me with helpful comments and feedback related to my research in these areas.

All the results provided in this dissertation were made possible by years of research collaboration with all my colleagues and the broader scientific community to which I am fortunate to belong. Especially, I am grateful to members of the ENCODE and IGVF consortia for building and generously sharing the expansive database of regulatory genomics that allowed some of the key findings detailed in this dissertation.

Lastly, I thank my family for their unconditional love and unfaltering belief in me.

Table of Contents

List of Figures

Chapter 2: Background

Chapter 3: Functional Characterization of Gene Regulatory Elements with CRISPR

Chapter 4: Modeling Human and Mouse Enhancer Sequences with Machi Learning

Chapter 5: Evolution and Conservation of Mammalian Enhancers

Chapter 6: gkm-align: an algorithm to map conserved distal enhancers using gapped-kmer sequence features

Chapter 7: Algorithmic details of gkm-align

Chapter 1

Introduction

1.1. Overview

The human genome encodes around 20,000 protein coding genes, and precise regulation of cell-specific gene expression is crucial for the development of diverse cell/tissues with distinct protein expression profiles and biological functions. Gene expression is modulated through non-coding regulatory DNA elements such as enhancers and promoters, and their mutation can cause disruption in gene regulation and lead to a wide range of common diseases such as cancer and schizophrenia[1]. Understanding functional and sequence properties of gene regulatory elements is indispensable for elucidating physiological processes driven by gene regulatory elements, such as embryonic development[2], and for identifying pathogenic regulatory mutations[3,4] that disrupt such processes. However, studying gene regulatory elements, especially enhancers, has proved difficult due to their complex contribution to gene expression and their degenerate sequence structures[5]. These properties of enhancers also induce rapid evolution of enhancers[6,7], which makes it difficult to functionally characterize enhancers through model animals such as mice. This dissertation is focused on developing and applying novel technologies to overcome such obstacles for studying enhancers and expand our knowledge of gene regulation. The dissertation is composed of two parts, with emphasis on the second part: 1) applying CRISPRi technology for direct functional characterization of diverse gene regulatory elements[8,9] and 2) developing novel genome-alignment algorithm[10] to map human distal enhancers to mouse using sequence features that effectively model the degenerate sequence structure of enhancers: gapped-kmers. A large portion of the dissertation will focus on the second part.

Enhancers contain degenerate clusters of binding sites for transcription factors (TFs)[5], and it is widely accepted that enhancers activate transcription through physical interaction with target promoters[11]. However, despite the consensus on their general role as a class of DNA elements, mapping individual enhancer to biological functions has been difficult. For example, identifying an enhancer's target promoter is not trivial as enhancers are often distal to their target promoters, often skipping multiple genes[12–14]. How the genome modulates interaction specificity among enhancers and promoters is one of the key active research areas. Also, many enhancers are accompanied by other enhancers that regulate the same target gene[8,15,16]. The enhancer redundancy may confer robustness against regulatory mutations or modulate target gene expression cooperatively or independently[11], but regulatory dynamics of most gene regulatory circuits encoded in the genome are yet to be characterized. All these obstacles against understanding regulatory programs have been largely due to the lack of experimental techniques that directly characterize enhancer functions in their native biological contexts. Since its invention, CRISPRi (dCas9-KRAB)[17], has become a popular technology for directly characterizing enhancer elements by delivering target specific regulatory perturbation in vivo. We collaborated with members of the ENCODE and IGVF consortia to systematically perturb diverse gene regulatory elements in cancer cells and embryonic stem cells to dissect regulatory circuits of multiple gene loci of significant biological interest and identified important principles of gene regulation[8,9]. The dissertation will briefly cover the results from the collaborations, with emphasis on computational analysis of enhancer perturbation.

Expanding functional characterization of enhancers beyond the level of gene expression or cellular phenotype requires testing their orthologous counterparts in model animals such as mice[2,13,15,18,19]. For example, to validate a putative pathogenic mutation in a human enhancer, an orthologous enhancer element in mice can be genetically

engineered to confirm whether the regulatory mutation in mice also leads to increased risk of the disease[2,15,19]. However, mapping conserved enhancers between distant mammals is difficult due to the rapid evolution of enhancers. Many enhancers contain multiple redundant and degenerate transcription factor binding sequences, and such sequence structure of enhancers tolerate accumulation of regulatory mutations with limited impact to enhancer activity[5,6]. Further, enhancer redundancy in gene regulation allows gain or loss of function mutations in enhancers with only mild impact to gene expression[16]. Such properties of enhancers allow rapid enhancer turnover, and as a result, many enhancers lack conserved markers of enhancer activities (e.g., chromatin accessibility) at orthologous loci identified by conventional genome alignment and mapping algorithms[20]. The lack of apparent regulatory conservation is largely due to the rapid evolution of enhancer elements, but limitations in conventional genome alignment algorithms[21,22] can also contribute significantly. Conventional methods align enhancers by detecting colinear sequence of nucleotides using variants of the Smith-Waterman algorithm, but such strategy may not be suitable to model the degenerate sequence structure of enhancers. Gapped-kmer sequence features effectively model the sequence degeneracy, and previous works[3,4,23,24] from our lab have shown that gapped-kmer sequence-based machine learning models can accurately distinguish enhancers from DNA elements with no regulatory activities. A major focus of my Ph.D. research has been to develop a novel genome alignment algorithm[10] that uses gapped-kmer based sequence similarity metric to identify optimal alignment paths along human and mouse genomes. From more than 2,000 human and mouse DNase-seq data from the ENCODE database, we derived 45 pairs of orthologous human and mouse cell/tissues with highly matching transcription factor activities. Using this expansive dataset, I show that *gkm-align* can identify more than 20,000 novel conserved enhancers across the 45 cell/tissues. Through the analysis of this expansive human and mouse data, we also observed intriguingly high level of cell/tissue

specific variation in enhancer conservation. The pattern of cell/tissue specific variation in enhancer is consistent with the previously reported pattern of cell/tissue specific conservation level of gene expression[25]. For example, conservation levels of both brain enhancers and gene expression are significantly higher than those of the liver and immune cells. This pattern of variability is largely explainable by association with transposable elements, which provides a quantitative support to the intriguing idea that transposable elements may play a central role in propagating TF binding sites and drive evolution of gene expression[26,27].

To expand our knowledge of the regulatory genome, we collaborated with multiple ENCODE and IGVF consortia labs to functionally characterize diverse gene regulatory elements[8,9] through systematic epigenomic perturbation with CRISPRi. Further, to facilitate future functional characterization of enhancers through model animals such as mice, we developed a novel genome alignment algorithm[10], through which we observed important principles of enhancer evolution. Together, I hope the results I present in this dissertation will be a valuable resource for future studies of gene regulation and spur development of treatments for genetic diseases caused by regulatory mutations.

1.2. Thesis organization

In **chapter 2**, I provide background information on enhancer biology. I review how interplays among enhancers, promoters, and regulatory proteins dynamically control gene expression. Then, I briefly go over some key biological properties of enhancers (e.g., chromatin accessibility) and standard experimental methods that are widely used for mapping putative enhancers genome-wide (e.g., DNase-seq). This information will serve as the scientific basis for all the analysis I will present in the subsequent chapters.

In **chapter 3**, I present the results from the research collaboration with members of the ENCODE and IGVF consortia, through which we systematically perturbed diverse gene regulatory elements to directly characterize their biological functions[8,9]. I first provide an overview of how CRISPR is utilized to functionally characterize many putative enhancer elements in parallel. Then, I present some of the key findings, from the collaboration with ENCODE, with emphasis on the optimal design of CRISPR screens and the strand-biasedness of CRISPRi perturbation at gene-bodies. Then, I will present computational modeling of the regulatory process that drives differentiation of embryonic stem cells to definitive endoderm, using CRISPR screen data generated in collaboration with Dr. Huangfu (IGVF) and her team.

All the remaining chapters detail results from Oh and Beer[10], discussing enhancer evolution and algorithms for detecting and characterizing conserved regulatory sequences.

Chapter 4 focuses on using a sequence-based machine learning method (gkm-SVM[3,23]) to model enhancer sequence structures across diverse human and mouse cell/tissue types[10]. This chapter introduces gapped-kmers, which is a sequence feature utilized by both gkm-SVM and the novel *gkm-align* algorithm (detailed in chapter 6). Then, I will show how gkm-SVM can be trained on enhancers to extract kmer-weight vectors that encode TF binding motifs enriched in enhancers, which I will call 'enhancer regulatory vocabulary' for convenience. Using simulated enhancer sequences with seeded TF binding motifs, I will demonstrate that gkm-SVM regulatory vocabularies can effectively and reliably recover expected sequence motifs. The simulation study, complemented with previous experimental evidence[3,28–30], provides justification for using gapped-kmer sequence features to model enhancer conservation (**chapter 6,7**). Lastly, I show that TF binding motifs in enhancers are highly conserved between human and mouse in a highly cell/tissue specific manner[10]. This allows derivation of 45 human and mouse cell/tissue

pairs with highly similar sets of core TF regulators, which I use to quantify enhancer conservation rate by cell type (**chapter 5**) and to evaluate gkm-align's performance (**chapter 6**).

Chapter 5 provides quantitative analysis of enhancer conservation[10] between human and mouse using the 45 cell/tissue pairs derived in chapter 4. I will first provide a brief overview of enhancer evolution. I show that enhancer conservation is highly variable across cell/tissue types despite the consistently high levels of conservation of enhancer regulatory vocabularies. Then, I show that the pattern of cell-specific rate of enhancer conservation is largely explainable by association with transposable elements. I argue that the results I discuss in chapter 5 provide a quantitative support for the idea that mammalian gene expression evolution is driven by transposable elements carrying human TF binding sites.

In **chapter 6**, I evaluate gkm-align (algorithmic details in chapter 7) against conventional genome alignment and mapping algorithms. Further, I show that gkm-SVM regulatory vocabularies can be incorporated into gkm-align to maximize the number of identifiable orthologous enhancers conserved between human and mouse. Further, enhancer vocabularies and gapped-kmer sequence similarity metric can be combined to predict functional conservation of orthologous enhancer pairs identified with gkm-align, based on likelihood of functional conservation[10]. Utilizing this novel method, we generated an expanded and quantitatively ranked catalogue of conserved human and mouse enhancers across diverse tissues, I expect this will facilitate future discovery and functional characterization by prioritizing enhancers for testing in model animals. **Chapter 7** focuses on the algorithmic details of gkm-align[10].

Chapter 2

Background

The human body is comprised of trillions of cells with mostly identical genomic DNA sequences. Each of the genomic copies of the DNA contains around 20,000 protein coding sequences, and each of the trillions of cells selectively expresses a subset of the protein coding genes at variable degrees for their specialized cellular functions. The ability to regulate cell-specific gene expression is essential to multicellular organisms, and elucidating the mechanism of gene regulation has been one of the major scientific pursuits over the past century. Only around one percent of the human genome encodes protein coding sequences, and the rest of the 99 percent of the genome, the non-coding genome, has mostly unclear biological functions. Much of the non-coding genome may be non-functional and neutrally evolving remnants of the past evolutionary history, but it has become clear through the past decades of research that an important fraction of the non-coding genome encodes instructions for regulating gene expression[2,31,32]. This information is written in gene regulatory elements[1] such as enhancers and promoters. In this chapter, I give a brief overview of the biology of gene regulation, and provide some canonical features of gene regulatory elements that have allows genome-wide mapping of candidate gene regulatory elements. This information will serve as the basis for functional (**chapter 3**), sequence (**chapter 4**), and evolutionary characterization (**chapter 5-7**) of enhancers in subsequent chapters.

[1] In this dissertation, I will use the terms 'gene regulatory elements' and 'cis-regulatory elements (CRE)' interchangeably.

2.1. A brief overview of gene regulation

Gene regulation is modulated through interplay among gene regulatory DNA elements (enhancers and promoters) and regulatory proteins (transcription factors and histones). Transcription factors (TFs) are regulatory proteins that directly bind to DNA with binding preference encoded in their amino acid sequences. Enhancers and promoters contain clusters of TF binding motifs, which determine their binding affinity and specificity to TFs. Promoters, located near transcription start sites (TSS), are known to bind basal transcriptional machinery such as RNA polymerase II (PolII) and general transcription factors which are ubiquitously expressed across cell types. On the other hand, enhancers are often distal to TSS and bind transcription factors with cell-specific expression[24]. Enhancers activate with increasing TF concentration and binding[33,34], which in turn recruit cofactors. Then, the recruited cofactors can mediate transcriptional machinery assembled at promoters[35] or remodel nucleosome stability for transcriptionally permissive chromatin environment[34,36] through direct long-range physical contacts[11].

Enhancers are modular and contain clusters of TF binding sites (TFBS) with flexible DNA sequences, and such sequence properties confer enhancers functional robustness to many mutations. There are well-documented examples of eukaryotic enhancers with strictly ordered arrangement of TFBS [5,37–40], but these examples are likely to be exceptions. In many cases, enhancers appear to tolerate addition or loss of TFBSs or change in their positions, causing little phenotypic impact in vivo and potentially inducing frequent TFBS turnover[41–44]. Divergent enhancer sequence architectures with varying TFBS positions have shown indiscernible biological functions[45,46], and, therefore, strict TFBS positioning rules likely do not exist for most enhancers, and this is reflected in performance comparisons of machine learning methods at predicting reporter expression from enhancer sequences[4,28,47,48]. In these studies, a method that represents enhancers

as position-independent clusters of TFBS (gkm-SVM[3,23]; more detail in **Chapter 4**) had performance comparable to, if not better than, methods that incorporate TFBS positions and orders using convolutional neural networks[49,50]. This suggests that arrangement of TF binding sites is not an essential parameter that determines function for most enhancers, although some level of modulatory effects probably does exist. Combined with the degeneracy of TF-DNA binding that tolerates a level of sequence variations within a TFBS[51,52], the disordered and modular TFBS positioning of enhancers allows design flexibility at low fitness costs, facilitating dynamic evolution of enhancers. This sequence property of enhancers is central to building the *gkm-align* algorithm[10], which measures enhancer conservation by comparing gapped-kmer sequence compositions between human and mouse sequences (algorithmic detail in **Chapter 7**).

Enhancer activation is modulated through DNA binding competition between TFs and histones. Due to the electrostatic attraction between the negatively charged DNA backbone and positively charged histone lysine residues, much of the eukaryotic genome wraps around histones, acting as a roadblock to TF binding at enhancers[34]. Therefore, activation of an enhancer requires a high local concentration of TFs with sufficiently high affinity to the enhancer, enough to compete away the histones bound to the enhancer. Such competition among DNA binding proteins is persistent across almost every segment of the genome[53,54]. At the level of a single enhancer, an enhancer can be considered a signal integrator, having TF concentrations as inputs and its activation as the output. It is curious what regulatory role histones play at the level of an enhancer other than blocking TF signal input with steric hindrance, and this question is elegantly addressed with the model of nucleosome-mediated TF cooperation described in Mirny[55]. This model shows that, even without any direct TF-TF physical interaction, presence of multiple TF binding sites at an enhancer leads to cooperative TF binding and competition against histones,

and this allows sigmoidal enhancer activation with respect to the input TF concentration (**Figure 2.1A**; reproduced using equation [1] and [2] from Mirny[55]). This mathematical model shows that the nonlinear sigmoidal response of enhancers requires multiple TF binding sites and nucleosome stability, which are parameters tunable with natural genetic (enhancer mutation) and epigenetic changes (enhancer methylation, histone remodeling) or artificially with CRISPR (detailed in **Chapter 3**). With the sigmoidal activation, a small change in TF concentration at low TF concentration leads to minimal change in enhancer activation. On the other hand, with low DNA-histone affinity (red line in **Figure 2.1B**) or with only a single TFBS (red line in **Figure 2.1C**), enhancer activation linearly increases with TF concentration in the low TF concentration limit. Since local TF concentration can stochastically fluctuate, such lack of sigmoidal response can lead to leaky transcriptional activation and perturb cellular phenotypes. Hence, enhancers typically have combinations of multiple TFBS, which leads to a robust biological switch that discriminately responds to changing TF concentration.

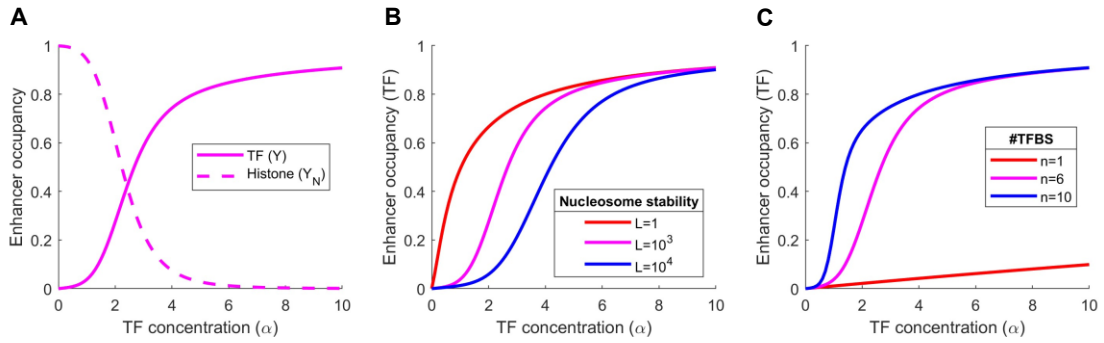

Figure 2.1. Nucleosome mediated activation of enhancers with varying transcription factor (TF) concentration

Reproduced using equation [1] and [2] from Mirny 2010. **A)** Enhancer occupancy by TF (solid line; $Y = \alpha \cdot \frac{(1+\alpha)^{n-1}+Lc(1+c\alpha)^{n-1}}{(1+\alpha)^{n}+L(1+c\alpha)^{n}}$) and histone (dotted line; $Y_N = \frac{L(1+c\alpha)^{n}}{(1+\alpha)^{n}+L(1+c\alpha)^{n}}$) with varying TF concentration (α); n=6, L=10^3, c= 0.01. **B)** TF enhancer occupancy at low (L=1), intermediate (L=10^3), and high (L=10^4) nucleosome stability (L); n=6, c=0.01 **C)** TF enhancer occupancy with n=1, 6, and 10 TF binding sites within the enhancer; L=10^3, c=0.01

Upon activation by TFs, enhancers can then regulate gene expression through distal interaction with target promoters, often redundantly with other enhancers[11]. The most well-studied example of distal enhancers is perhaps the ZRS enhancer that regulates the expression of the SHH gene in the limb bud during embryonic development[12]. Mutations of the ZRS enhancer are known to cause developmental defects such as polydactyly through disruption of Shh expression[2]. The ZRS enhancer is almost 900 kilobases away from the Shh gene, regulating Shh expression through direct physical contact. Contact frequency between enhancers and promoters tend to correlate with gene regulation[19,56], and accumulating evidence suggests that convergent CTCF binding motifs recruit CTCF and cohesion architectural proteins to shape the specificity of enhancer-promoter interactions (EPI) through formation of topologically associated domains (TADs)[8,11,57–64]. TADs are units of genomic compartments (~1Mb) segregated by high pairwise DNA contact frequency[11]. The CTCF-driven three-dimensional chromatin organization induces complex EPI networks, where EPI frequency sharply decreases across TAD boundaries, while decreasing relatively mildly with genomic distance within TADs[56,65]. For example, mouse embryos that were genetically modified to have increased distance between ZRS and Shh without TAD disruption experienced only a mild decrease in Shh expression and retained normal limb development. On the other hand, mouse embryos with structural variations that disrupted the TAD boundaries led to monodactyly, even when the structural variations brought ZRS and Shh closer[19]. These observations suggest that distal enhancers can be variably positioned within TADs with mild regulatory impact. As shown in **Chapter 5**, CTCF binding sites are highly conserved between human and mouse with little variation by cell-type, suggesting that TADs are evolutionarily and developmentally invariant regulatory compartments. Enhancer redundancy further adds to the design flexibility of EPI networks. Many mammalian promoters have been shown to physically interact with multiple enhancers[11,15,66,67], where the number of interacting

enhancers correlate with the level of downstream gene expression[11,66,67]. Enhancer redundancy provides protection against phenotypic perturbation from loss-of-function mutations in individual enhancers[8,15], and it can modulate gene expression by additively increasing the overall time that a promoter is in contact with any one of the redundant enhancers[8,11].

Together, the complex network of gene regulatory DNA and protein elements orchestrates gene expression, and cell-specific expression of TFs define cell-states. The increase in TF concentration leads to selective activation of enhancers containing associated TFBS, modulable with flanking histone modifications; activated enhancers then recruit cofactors to distally regulate gene expression of their CTCF-defined insulated neighbors. This picture of gene regulatory landscape can be visualized as **Fig 2.2**. This principle of gene regulation is highly conserved across mammals, while the design flexibility of the network structure facilitates mammalian evolution through transcriptional rewiring. This will be discussed in more depth in **chapter 5**.

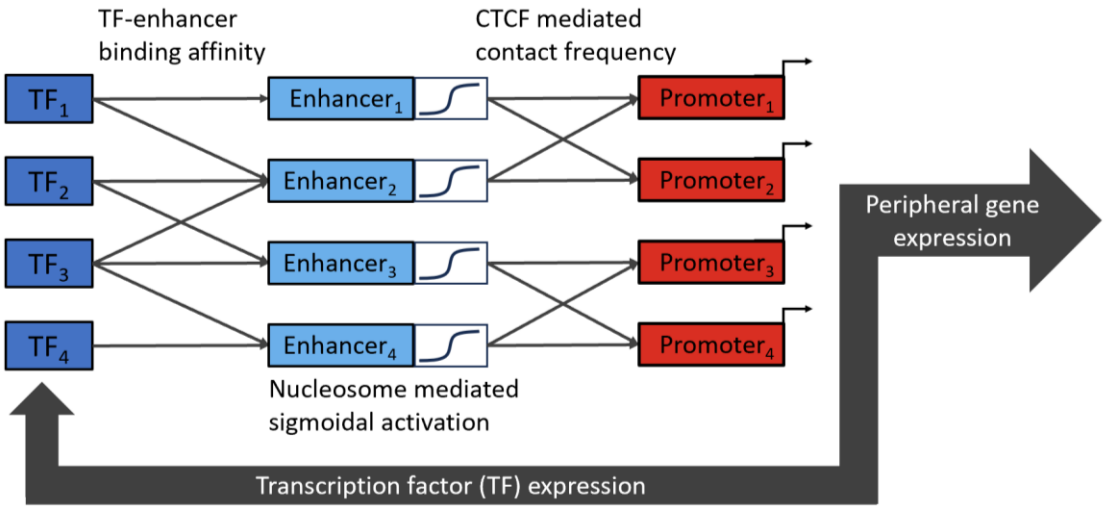

Figure 2.2. Transcriptional network wiring for regulation of gene expression.

Binding affinity between transcription factors (TF) and enhancers modulates their connectivity. Distal interaction mediated by CTCF-loops modulate connectivity between enhancers and promoters. Downstream gene expression determines cellular functions and identities. TF expression feeds back to the regulatory network.

2.2 Genome-wide identification of putative gene regulatory elements

As discussed in the previous section, gene regulatory elements are chromatin accessible, marked by characteristic histone modifications at flanking nucleosomes, and contain TFBS under evolutionary selection pressure. These properties of gene regulatory elements were initially generalized from individual characterizations of a relatively small number of gene loci using low-throughput experimental technologies[68–76]. These techniques soon transformed into high-throughput genome-wide methods[77–80] following advancements in sequencing technologies, and have spurred large collaborative projects[32,81,82] for the ambitious goal of mapping every regulatory element across the human genome for diverse cell types. In this section, I will review some of the key high-throughput experimental methods for mapping gene regulatory elements (More comprehensive reviews in Maston et al.[83], Klemm et al.[34], and Gasperini et al.[84]). All these methods involve purifying DNA fragments with biological properties associated with regulatory activity and mapping to a reference genome for signal enrichment.

ChIP-seq

Chromatin immunoprecipitation (ChIP)-seq is widely used to map DNA loci bound by specific regulatory proteins of interest. A typical ChIP-seq procedure[83,85] begins with cross-linking cells with formaldehyde, which creates covalent bonds between interacting nucleic acids and proteins, followed by sonication to shear the DNA into fragments of 100-300 base pairs. Then, the DNA fragments are purified using antibodies with high binding specificities to the protein of interest. After reversal of cross-links, the enriched DNA fragments are sequenced and mapped to a reference genome using read-mapping software[86,87]. The mapped sequencing reads can then be piped into peak calling software[88] to identify putative regulatory elements with significant binding association with the protein of interest.

One major challenge of utilizing ChIP-seq for enhancer mapping is that it requires prior knowledge of molecular markers that define enhancers. Early studies have shown that histone acetyltransferases, such as p300, are associated with active enhancers and promoters[89,90]. Mapping p300 binding sites with ChIP demonstrated that active enhancers are further associated with flanking histone modifications such as mono-methylation of histone H3 lysine 4 (H3K4me1) and acetylation of histone H3 lysine 27 (H3K27ac)[91]. Promoters are also enriched with H3K27ac, but marked by tri-methylation of histone H3 lysine 4 (H3K4me3) instead of H3K4me1. Genome-wide analysis has shown the markers of active enhancers are highly cell-specific and correlate with expression of nearby genes[92], while histone modifications at promoters appear cell-type invariant. These markers of enhancers and promoters are highly conserved, and have been extensively utilized to map putative regulatory elements across diverse mammals[20,93]. However, identifying enhancers using histone marks has limitations. First, histone marks tend to be broad, often extending several kilo bases, and make it difficult to identify core TF binding sequences[83]. More importantly, although histone marks associated with regulatory elements are well-defined, the regulatory mechanism of how each histone modifications play are not yet clear and still under active research[94]. ChIP-seq signal of TF binding is relatively narrower and allows more precise identification of enhancer coordinates. However, identification of enhancers using TF ChIP-seq requires prior knowledge of active TFs for a given cell-type. Enhancer-associated TFs tend to have high cell-specific expression, and identification of active TFs is often difficult as TF expression levels tend to be smaller compared to the level of peripheral gene expression. Hence, the utility of ChIP-seq in enhancer identification is currently limited largely by our knowledge of mechanisms of gene regulation and structures of cell-specific gene regulatory networks.

DNase/ATAC-seq

As discussed in **chapter 2.1**, DNA-binding competition between histones and TFs determines activation of regulatory elements, where increasing TF concentration leads to higher TF occupancy and decreased histone occupancy at enhancer centers. As TF bindings tend to be unstable and have relatively lower residence time than histones[95,96], increased TF occupancy allows higher chromatic accessibility to enzymatic cleavage by DNase-I and Tn5 transposase through stochastic molecular interactions[34]. This is the molecular principle behind DNase-seq (DNase I hypersensitive site sequencing)[78,97,98] and ATAC-seq (Assay for transposase-accessible chromatin using sequencing) [99]. DNase-seq involves cleaving DNA sites with high chromatin accessibility with DNase I endonuclease. The fragmented DNAs can be filtered for short lengths to enrich for fragments that fall within regulatory elements[98], followed by sequencing and mapping to reference genome to identify DNase I hypersensitive sites (DHSs). Similarly, ATAC-seq involves cleaving accessible DNA sites with Tn5 transposase. Tn5 transposase further adds sequencing adapters to the cleavage sites, which are then used to sequence and map Tn5-accessible DNA fragments to a reference genome. Chromatin accessibility measured by DNase-seq and ATAC-seq are highly correlated[34]. Further, chromatin accessibility exhibits high correlation with aggregate TF ChIP-seq signals[53] (R = 0.79; sum of 42 TF ChIP-seq in K562 cells), supporting that chromatin accessibility is driven by TF binding. Also, consistent with the observations made with enhancer and promoter associated histone marks[92], promoters tend to be constitutively accessible across many cell-types, while distal enhancers tend to have cell-type accessibility[34,53]. As will be shown in **chapter 4**, distal DHSs with cell-specific accessibility are enriched with histone modifications associated with enhancer activities (H3K4me1, H3K27ac) and contain cell-specific TFBS sequence motifs, while proximal DHSs with constitutive accessibility are enriched with

promoter histone marks (H3K4me3, H3K27ac) and lack cell-specific TFBS sequence motifs. Hence, analysis of chromatin accessibility, histone modifications, TF binding, and underlying DNA sequence together provide a consistent and comprehensive portrait of the mammalian regulatory landscape.

2.3 Discussion

In this chapter, I introduced key properties of enhancers and discussed methods to map enhancers genome-wide. This information will serve as the basis for the remainder of the dissertation. Background information on other topics, such as enhancer conservation and algorithms for mapping conserved regulatory elements, will be provided in the beginning of each relevant chapter.

Chapter 3

Functional Characterization of Gene Regulatory Elements with CRISPR

Gene regulatory elements are identifiable using biomarkers such as chromatin accessibility, TF binding, and histone modifications. During the past two decades, millions of putative cis-regulatory elements (CREs) across diverse human and mouse cell/tissues have been identified through large consortium collaborations, such as the ENCODE consortium[32,81,100]. CREs are highly enriched with common pathogenic variants[1], and yet biological functions of most CREs have not been characterized. For example, gene targets for most putative enhancer elements identified with DNase-seq and TF/histone ChIP-seq assays are largely unknown. This critical missing piece precludes linking regulatory mutations to disruption in gene expression and to downstream pathogenesis. However, with recent development of CRISPR epigenome engineering techniques[101,102], functional characterization of CREs has become substantially more feasible. CRISPR allows direct and targeted perturbation of CREs, and the impact of CRISPR perturbation can be measured at the level of both gene expression[103,104] or phenotype[8,105]. In this chapter, I will first introduce background information on CRISPR and different types of CRISPR screens for direct functional characterization of CREs (**chapter 3.1**). Then, I will discuss two research collaborations, through which my collaborators and I utilized CRISPR to functionally dissect CREs of high biological interest. The first collaboration[2] to be

[2] This work was completed with equal contribution from Drs. Steve Reily, David Yao, Josh Tycko, Lexi Bounds, Sager Gosai, Lazaros Lataniotis and, myself, with additional contributions from many other members of the consortium. In this dissertation, I will specifically focus on analysis I contributed, which was conducted through close collaboration and discussion with all the main authors of the study, particularly Drs. Reilly (the cell

discussed involves systematic comparison and quantitative analysis of diverse CRISPR screens performed by the ENCODE consortium[9] (**chapter 3.3.1**). Next, I will discuss the collaboration[3] with Dr. Danwei Huangfu and her team (**chapter 3.3.2**); we CRISPR screened enhancers of core transcription factors that drive differentiation of embryonic stem cells (ESC) to definitive endoderm (DE)[8]. Most of the results presented in this section are adapted from the original publications of these studies[8,9], modified to be aligned with the goal of the dissertation.

3.1 Introduction to CRISPR-based functional characterization of CREs

A common strategy for dissecting a complex system is to systematically perturb its components and quantify the effect to the system. Such strategy has been difficult to utilize in molecular biology due to technical limitation, and thus biological functions of DNA elements could only be indirectly inferred through statistical analysis linking phenotypes to naturally occurring genetic variants[2,106] or randomly induced mutations[107]. The discoveries of site-specific endonucleases[108] and the inventions of gene editing tools such as ZFN[109] (zinc-finger nuclease) and TALEN[110] (transcription activator-like effector nuclease) opened up the possibility of targeted perturbation of the genome[111]; however, DNA-binding specificities of these tools are encoded in their DNA-interacting protein domains, and manipulating the DNA-recognition domain to target a specific DNA element in the genome is difficult. Upon its discovery, CRISPR was quickly adapted as a

coverage analysis) and Tycko (the strand-bias analysis). For a comprehensive discussion of the study, please refer to Yao et al[9].

[3] All the experiments in this study were performed by Dr. Danwei Huangfu and her lab members, with major experimental contributions from Drs. Renhe Luo and Jielin Yan. The computational models in this study were developed with a major contribution from my advisor, Dr. Michael Beer, with contributions from our lab members Drs. Dustin Shigaki, Wang Xi, and myself. In this dissertation, I will limit the discussion to only include computational analysis I contributed. For example, the CIA (CTCF interaction) model, developed exclusively by Drs. Beer and Xi, is not discussed. For a comprehensive discussion of the study, please refer to Luo et al.[8].

mainstream molecular tool for gene editing for this reason; the CRISPR system uses RNA, complementary to its target DNA sequence, to guide the effector protein to the target site[101]. Due to its ease of use and versatility, CRISPR has been successfully utilized to systematically perturb and characterize protein-coding genes[102,112,113] and gene regulatory elements[8,103–105,114].

CRISPR-Cas, in its natural state, is a prokaryotic adaptive immune system used against viral infection[115]. One of the early evidences of CRISPR (clustered regularly interspaced short palindromic repeats) came from a comparative genomic analysis that identified the CRISPR locus as a hypervariable polymorphic region unique to prokaryotes and absent in eukaryotes and viruses[116]. The CRISPR locus in prokaryotes are typically composed of multiple palindromic repeats (around 30 bps in length) separated by spacers, accompanied by nearby CRISPR-associated (Cas) genes[115]. The number of repeats and spacers was found to be highly variable across bacterial strains[117], where spacer sequences are often homologous to viral sequences[118]. This led to the hypothesis that CRISPR-Cas is a bacterial adaptive immune system against viruses. The hypothesis was soon confirmed through experiments that demonstrated that phage infection leads to addition of new spacers derived from phage genomic sequences and that the addition or removal of spacers leads to variable immune response against phages[115]. Further, inactivation of Cas genes led to inactivation of phase resistance, establishing Cas proteins as the essential components for CRISPR adaptive immunity[115]. Studies of the CRISPR system across diverse prokaryotes led to discoveries of diverse versions of the CRISPR-Cas system, with a large variety of Cas proteins with distinct functions and diverse mechanisms of actions[101,119].

The discovery of the CRISPR-Cas system led to the development of artificial CRISPR-Cas systems engineered for target-specific edition of mammalian genomes[120].

The most widely used version is composed of a chimeric single guide RNA (sgRNA)[121] and Cas9 protein, which cause double stranded breaks at target DNA loci, complementary to the sgRNA protospacer sequence (~20 bp long) adjacent to a PAM sequence (protospacer adjacent motif; 'NGG' for Cas9). In eukaryotes, double-strand breaks (DSB) are usually repaired through nonhomologous end joining (NHEJ), which often leads to small insertions or deletions at CRISPR target sites. Through this mechanism, targeting exons with Cas9 causes gene inactivation, and this strategy has been successfully employed to screen gene functions genome-wide[113,122]. However, one disadvantage of using Cas9 for functional characterization of DNA elements is that off-target Cas9 delivery can cause DSB-induced cytotoxicity, confounding downstream analysis. Instead, for functional characterization of CREs, Cas9 proteins with deactivated endonuclease domains (dCas9) can be delivered to sterically block TF binding[123]. For higher repressive potential, repressive epigenetic regulators, such as KRAB[124] (Krüppel-associated box), can be fused to dCas9[17], and this modality of CRISPR perturbation is commonly referred to as CRISPRi. The addition of the KRAB domain to dCas9 has been shown to drastically repress the expression of target genes[17] and affect downstream phenotypes[125]. This property of CRISPRi allows direct functional characterization of CREs through targeted repression.

CRISPRi, combined with next-generation sequencing, allows simultaneous functional screening of hundreds of CREs through parallel epigenetic perturbation. The first large scale CRISPR screen of CREs with CRISPRi was proposed and performed by Gilbert and colleagues[114]. In the study, they generated a K562 human myeloid leukemia cell line with stable dCas9-KRAB expression through viral transduction, and designed a large library of sgRNAs targeting the promoter regions of 15,977 protein-coding genes (10 guides per TSS). Each sgRNA was packaged into a lentivirus, and the lentiviral sgRNA

packages were delivered to a large population of dcas9-KRAB expressing K562 cells. The experiment was designed so that most of the cells received one or less sgRNA genomic insertion and that each unique sgRNA was delivered to ~4,000 cells (i.e. MOI<1, cell-coverage=4,000). The K562 cells, expressing dcas9-KRAB and promoter-targeting sgRNA, were grown for 10 days, followed by PCR amplification and sequencing of the sgRNAs. sgRNAs targeting the promoter of GATA1, a known TF activator of K562 proliferation, were depleted in the sequence reads after 10 days of growth competition, indicating that GATA1 promoter is a key CRE for cell growth. Such pooled CRISPR strategy allows unbiased and quantitative functional assessment of CREs, by generating diverse genetic/epigenetic variations and selecting cells based on downstream perturbation of phenotypes. The major strength of pooled noncoding CRISPR screen is that the experimental procedure can be flexibly modified to evaluate CREs for various biological functions, and variations of noncoding CRISPR screen strategy have soon been developed. For example, to identify enhancers that regulate a specific gene of interest, cells, infected with different sgRNA targeting putative enhancers, can be selected by expression of the gene. Fulco and colleagues combined CRISPRi with fluorescence in situ hybridization (FISH), targeting specific RNA molecules, and flow cytometry to sort sgRNA-infected cells based on expression of the gene (CRISPRi-FlowFISH[104]), and identified CREs of GATA1 and HDAC6, among other genes. Similarly, Reily and colleagues used hybridization chain reaction (HCR) FISH to amplify detection of gene expression, followed by fluorescence-based cell sorting (CRISPRi HCR-FlowFISH[103]). Lastly, sgRNA-infected cells can also be sorted by protein expression, for example, by tagging a gene of interest with a fluorescent protein (e.g. GFP)[8,126,127]. As will be discussed in subsequent sections, the development of CRISPR-based technologies has allowed functional dissection of CREs, which will continue to enhance our understanding of gene regulation and diseases caused by aberrant regulation of gene expression.

3.2 Methods

3.2.1 Identifying common sgRNA perturbations across ENCODE CRISPR screens at GATA1 locus

The GATA1 CRISPR guide quantification files were downloaded from the ENCODE portal (encodeproject.org), from which the sgRNA protospacer coordinates were extracted. Hg38 coordinates were used to uniformly analyze and compare the five CRISPR screens from various labs. For screens with hg19 coordinates, their protospacer coordinates were first mapped to hg38 using bowtie1 with "-n --best" option. Then, the hg38 PAM coordinates for each screen were extracted by taking the three base pairs downstream of each protospacer, which were confirmed to contain the expected NGG sequence. For the GATA1 locus, 250 such PAM coordinates were found to be shared across the five screens, and these common PAM coordinates were filtered out using sgRNA GuideScan-aggregated CFD specificity score [128,129] (>0.2), leading to 176 remaining PAM coordinates.

3.2.2 Computing sgRNA perturbation effect size from ENCODE *guide-quantification* files.

ENCODE guide-quantification files contain a row for each unique sgRNA delivered to the cells, each containing the number of sgRNAs integrated to the cell population. The numbers were quantified by mapping PCR amplified sgRNA sequence reads to the sgRNA library. For each of the CRISPR screens used in this study, there are two guide-quantification files, one for each of the two cell populations in comparison for sgRNA population.

For the expression-sorting based CRISPR screens (FlowFISH, HCR-FlowFISH), sgRNA infected cells were sorted into two populations by fluorescence, having low-expression or high-expression of the target gene. For growth screens, there are two cell populations: cells during the early and late time point of cell culture post sgRNA infection. For the analysis presented in this dissertation, we used cells cultured for 21 days post infection for the late-time-point cells and pre-infection plasmid pool of sgRNAs for the early- time-point (equivalent to day 0 post infection).

Denoting the number of the i^{th} sgRNAs mapped from the low-expression sorted or the early time point cells as $N_{A,i}$ and denoting the number of the i^{th} sgRNAs mapped from the high-expression sorted or the late time point cells as $N_{B,i}$, sgRNA perturbation effect size ($\log_2$FC) of the i^{th}sgRNA can be computed as:

$$\log_2\left(\frac{1+N_{A,i}}{1+N_{B,i}}\right)$$

That is, sgRNAs that lead to low expression of the target gene or lead to depletion through growth competition were assigned with high CRISPR perturbation effect size.

Log_2FC can be normalized using Z-transform or can be mean-normalized to underweight sgRNAs with low numbers of mapped reads. The mean normalization appears to generate more reproducible perturbation effect sizes (higher bioreplicate Pearson correlation[9]), especially for the sorting screens. The mean normalization was computed as:

$$\log_2\left(\frac{1+\frac{N_{A,i}}{mean(N_A)}}{1+\frac{N_{B,i}}{mean(N_B)}}\right)$$

3.2.3 Identifying exon or DHS targeting sgRNAs with significant perturbation effect size

sgRNAs were classified as exon-targeting if their PAM coordinates overlapped with Ensembl-annotated exons or as DHS-targeting if they overlapped with K562 DHS, where the DHS coordinates were obtained by extending K562 DHS narrow peaks (ENCODE accession ID: ENCFF899KXH) by 500 base pairs in both directions from their centers.

3.2.4 Sampling analysis for simulating CRISPR screens performed at various sequencing depths

For this analysis, we used ENCODE guide quantification files (described in **Methods 3.2.2**). To simulate an experiment with sequencing depth of d, we sampled with replacement total N×d number of reads independently from cell population A and B, where N is the number of distinct sgRNAs in the library. For the CRISPR screens used for the bio-replicate reproducibility and dropout analyses, reads were sampled independently for each of the two bio-replicates (A1, A2, B1, B2). sgRNAs that had 0 mapped reads in any one of A1, B1, A2, and B2 were excluded from the analyses. At each sequencing depth d, 100 independent bootstrap samples were generated to be used for the dropout and bio-replicate reproducibility analyses. For the dropout simulation analysis, we defined dropout sgRNAs as those that resulted in less than 10 sampled reads from either A or B.

3.2.5 Strand specific quantification of sgRNA effect sizes

All the HCR-FlowFISH CRISPR screens used in this analysis have specific gene targets (CRISPRi growth screen tiling across GATA1 locus, and HCR-FlowFISH), and their sgRNAs were unambiguously labeled as either template or coding strand targeting sgRNAs depending on which strand their protospacers are located relative to the

transcriptional directions of their target genes. For the GATA1 CRISPRi growth screen, sgRNAs were filtered for GuideScan-aggregated CFD specificity score >0.2 to remove sgRNAs with off-target growth effects. Then, we labeled each sgRNA as gene-targeting if its PAM sequence was located between 2,000 base pairs downstream of TSS and TES. The 2,000 spacers were used to exclude gene-body targeting sgRNAs that are TSS-proximal and affect promoter activities. sgRNAs with PAM sequences located between 2,000 base pairs upstream of the TSS and the TSS itself were labeled promoter-targeting, and all other sgRNAs were labeled as “outside”. Refgene annotations were used to identify TSS and TES for each gene, and for genes with multiple isoforms, isoforms with the highest levels of K562 Pol-II Chip-seq signals (ENCFF914WIS, signal P values) at both the TSS and TES were used. 3 out of 20 HCR-FlowFISH experiments were excluded for this analysis as they had less than 5 tested protospacers located within template-strand promoters, coding-strand promoters, template-strand gene bodies, or coding-strand gene bodies.

3.3 Results

3.3.1 Multi-center integrated analysis of noncoding CRISPRi screens (ENCODE)

The initial drafts of the human genome revealed that only around 1% of the genome codes protein-coding genes[130,131], with largely unknown biological function for the rest of the genome. Elucidating the role of the non-coding genome has been a major focus of the ENCODE (Encyclopedia of DNA Elements) consortium[32,82]. Over the past two decades, the ENCODE consortium has mapped around 1 million human and mouse putative CREs with biomarkers associated with regulatory activities. During its last phase, the ENCODE Functional Characterization Centers (FCC) CRISPR-screened a fraction of

the putative CREs. Overall, 107 CRISPR screens, comprising of >540,000 perturbations across 25 megabases of the genome, were performed by multiple FCC labs using various pooled CRISPR screen methods[103–105,132]. As a member the ENCODE consortium, I collaborated with the FCC labs 1) to set up the largest public repository of noncoding CRISPR screens to date (accessible through encodeproject.org) and 2) to perform meta-analysis of the CRISPR screen data through uniform computational pipelines. Using the large database of CRISPR screens, we developed experimental and computational guidelines for performing and analyzing CRISPR screens and discovered previously unreported properties of CRISPR perturbation that will have significant implication for designing future CRISPR screens of CREs.

Effects of common sgRNA perturbations demonstrate CRISPRi screens are reproducible.

We analyzed diverse CRISPR screen methods applied to K562, primarily focused on the GATA1 locus. GATA1 is a transcription factor known to promote cell proliferation, and has three well-characterized CREs in K562 cells (GATA1 promoter and two enhancers referred to as *e-GATA1* and *e-HDAC6* by Fulco et al[105]), making it an ideal target for systematic analysis of the CRISPR screens. For this locus, we analyzed and compared five different CRISPR screens: CRISPRi FlowFISH[104], CRISPRi HCR-FlowFISH[103], CRISPRi growth screen (from two labs[105,132]) and CRISPR-Cas9 growth screen[132] (**Figure 3.1A**). Each of the CRISPR screens has two to three biological replicates. The five GATA1 CRISPR screens share 241 common perturbation targets, identifiable by downstream PAM sequences (NGG). We filtered out low-specificity sgRNAs with GuideScan CFD specificity scores[128,129] less than 0.2 to minimize confounding effects from off-target perturbation, leading to total 176 common CRISPR perturbation shared across the five screens (**Methods 3.2.1)**. Effect size of each of the 176 common perturbations were computed for each of the screens using sgRNA enrichment ($\log_2 FC$; **Methods 3.2.2**), and

the five screens and their biological replicates were pairwise compared using Pearson correlation across the 176 guide effect sizes (**Figure 3.1B**). All the five CRISPR screen approaches had high reproducibility, with high pairwise Pearson correlation between replicates (n=5; Pearson Correlation; min: 0.59, max: 0.90, mean: 0.77). Overall, we observed that CRISPR screen results were similar among experiments with shared perturbation modalities (dCas9-KRAB vs. Cas9) despite the distinct perturbation readouts used (growth vs. expression). Experiments using CRISPRi exhibited high correlation among themselves (n=36; Pearson Correlation min: 0.42, max: 0.90, mean: 0.56), while

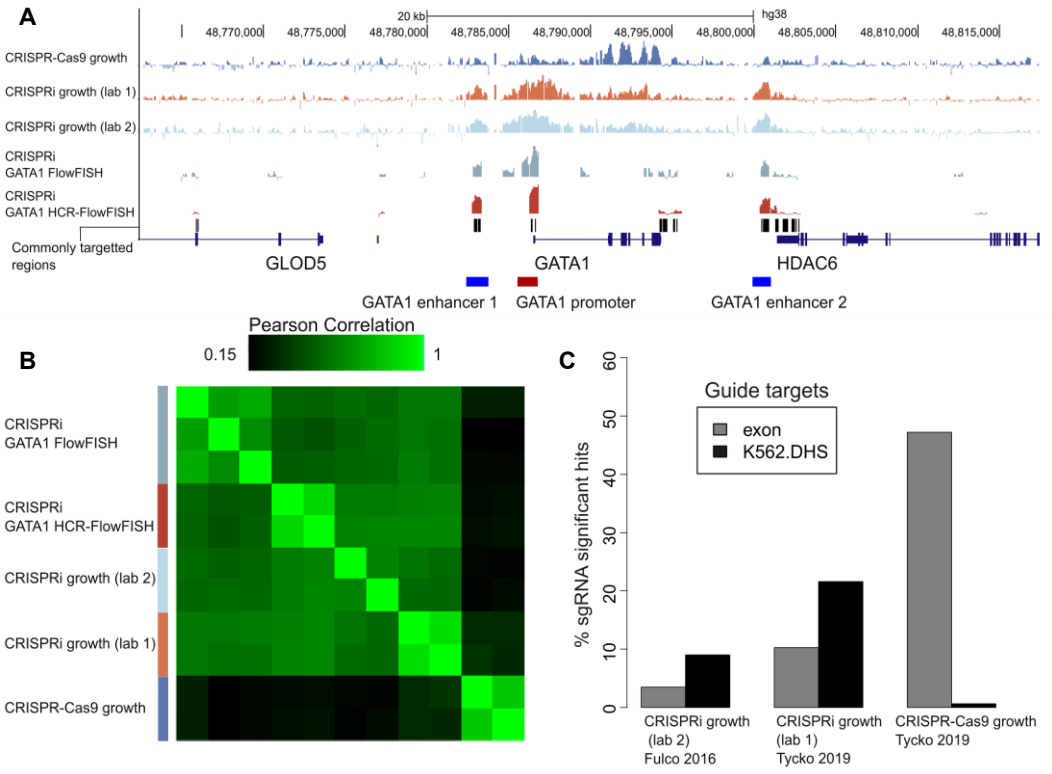

Figure 3.1. Effects of common sgRNA perturbations demonstrate CRISPRi screens are reproducible.

A) CRISPR perturbation signals (log2FC) at GATA1 locus **B)** Pairwise correlations (Pearson) of normalized effect sizes on sgRNAs commonly shared by each method (N=176 sgRNAs with GuideScan-aggregated CFD specificity score>0.2) in GATA1 locus; multiple rows for the same CRISPR method show replicates. **C)** Percentage of exon (grey, total N=3642, 1739, 1735 from left to right) or K562 DHS targeting guides (black, total N=3472, 1570, 1592 from left to right) with significantly high log2FC effect sizes (Z-score P value < 0.001)

the CRISPRi experiments had low correlation with the Cas9 growth screen low (n=18; Pearson Correlation min: 0.15, max: 0.32, mean: 0.21). The difference in effect sizes between CRISPR-Cas9 and CRISPRi screens comes from their mechanism of perturbation. We compared the two CRISPRi growth screens with the Cas9 growth screens, and observed that a high percentage of sgRNAs with significant perturbation effect is enriched in exon for Cas9 while higher percentage is enriched in CREs for CRISPRi (**Figure 3.1C; Methods 3.2.3**). We reasoned that Cas9 targeted at exons effectively knocks out GATA1 through the error-prone NHEJ repair process and inhibit K562 proliferation, while inducing heterochromatin state in GATA1 gene body with CRISPRi has much weaker effect at repressing GATA1's function. On the other hand, Cas9 has weaker effect at CREs likely because inducing small indels in CREs delivers weaker perturbation as CRE functions are often robust against indels or substitutions (discussed in **Chapter 2**). These results together point to the conclusion that CRISPRi reproducibly identifies CREs and that dCas9-KRAB is a more suitable choice of perturbation modality than Cas9 for functional characterization of CREs.

Cell coverage and sequencing depth impact CRE detection accuracy and sgRNA dropout

Then, we sought to generate an experimental guideline for performing CRISPR screens of CREs, through computational analysis of the CRISPR screens conducted by the ENCODE consortium. One of the key experimental parameters for CRISPR screens is the number of cells to which each unique sgRNAs is delivered: cell coverage. Another important parameter is sequencing depth, and we analyzed how cell coverage and sequencing depth affect the quality of CRISPR screens. To interrogate the effect of cell coverage, we quantified how accurately the sgRNA perturbations can recover the three

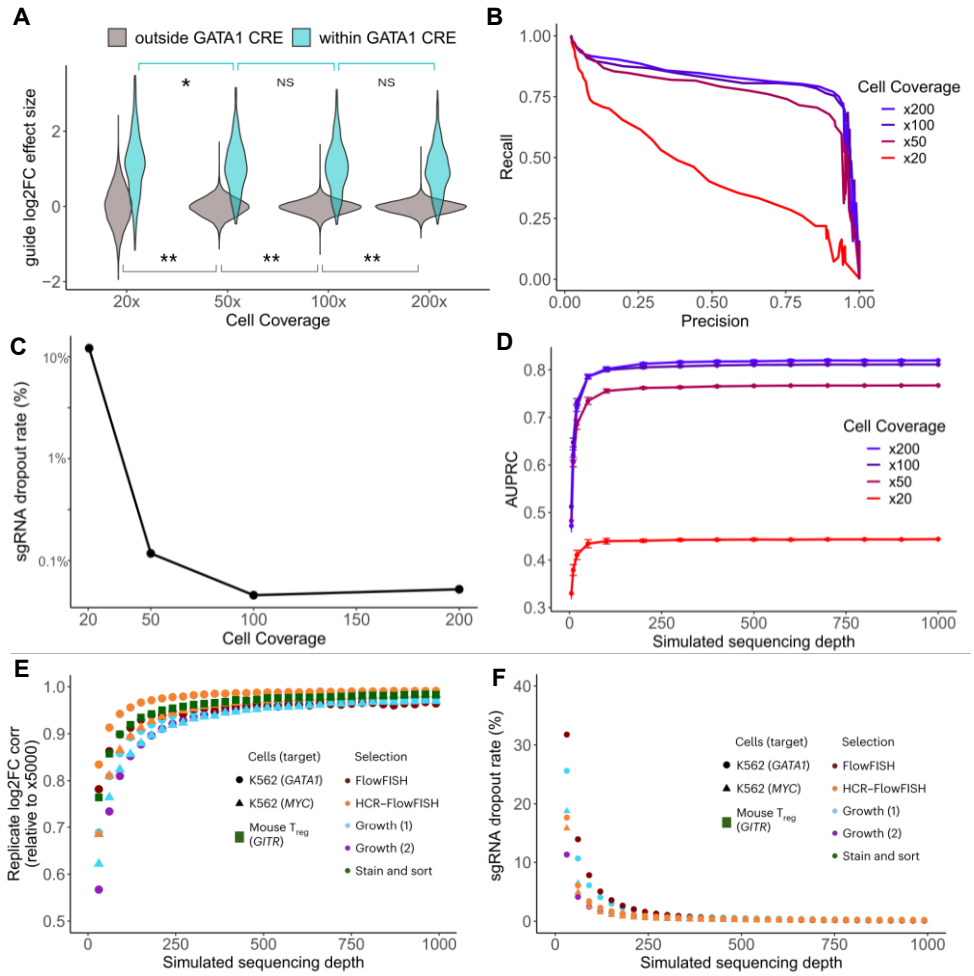

Figure 3.2. Cell coverage and sequencing depth impact CRE detection accuracy and sgRNA dropout

A) Distributions of HCR-FlowFish guidewise log_2FC effect sizes (total 13,732 PAMs targeted) at various cell coverages, separately for sgRNA targets within (N=288) and outside known *GATA1* CREs (N=13,444). Asterisk denotes significant change in variance (* is p≤0.01 and ** is p<2.2e-16 by Levene's test. n.s. is p>0.2). **B**) Precision-recall curve for identifying *GATA1* CRE-targeting sgRNAs using effect sizes from various cell coverages (AUPRC: 20x=0.44, 50x=0.77, 100x=0.81, 200x=0.82; CRISPRi HCR-FlowFish). **C)** sgRNA dropout rates (sgRNA with <10 mapped reads for low- or high-expression sorting bins) for varying cell sorting depth (20x, 50x, 100x, 200x) for CRISPRi *GATA1* HCR-FlowFISH performed at 2000x sequencing depth (n=1 replicate).**D**) AUPRC for identifying *GATA1* CRE-targeting sgRNAs with varying sequencing depth (bootstrap sampled) and cell coverages (20x, 50x, 100x, 200x). Dots and error bars indicate averages and 99% confidence intervals over 10 bootstrap samples. **E**) Biological replicate reproducibility (Pearson correlation of guidewise log_2FC), normalized to 5000x simulated sequencing depth and **F**) guide dropout rate (dropout defined as <10 mapped reads) in diverse CRISPRi screens with varying sequencing depth (bootstrap sampled). Dots show an average over 100 bootstrap samples. The GATA1 (circles) and MYC (triangles) screens in human K562 cells were performed with varied readout methods (colors); the GITR screen (rectangle) in mouse T-regs used protein staining followed by sorting. Growth datasets are (1) Tycko *et al.* 2019 and (2) Fulco *et al.* 2019.

known GATA1 CREs (GATA1 promoter, e-GATA1, e-HDAC6), at varying cell depths (**Figure 3.2A**).The CREs were also identified in all of our CRISPRi screens. CRISPRi HCR-FlowFISH was performed at cell coverages of 20x, 50x, 100x, and 200x (by Reilly and his team) with 288 sgRNAs targeting the three validated CREs and 13,444 sgRNAs targeting outside the three CREs. These screens were sequenced with sequencing depth higher than 2000x to further assess the impact of sequencing depth. Referring the CRE-targeting guides as *positive* and the other guides as *negative* sgRNAs, we computed the effect size of GATA1 perturbation for the positive and the negative sgRNAs, and tested whether the positive sgRNAs can be distinguished from the negative sgRNAs by effect size (**Figure 3.2A**). At a low cell coverage of 20x, the effect size of both positive and negative sgRNAs exhibited high variance as each sgRNA perturbation was applied to only a small number of cells, leading to low statistical support for effect size. As a result, the positive and negative sgRNAs were weakly separable by effect size, with low AUPRC of 0.44 (**Figure 3.2B**). With increasing cell coverage, variance of the negative sgRNA effect size approached zero with average effect size around zero. On the other hand, variance for the positive sgRNAs stabilized at cell coverage greater than or equal to 50x, reflecting the biological variance of the CRE strengths and efficiency of the CRE targeting guides. As a result, increasing cell coverage led to higher precision and sensitivity for distinguishing *positive* from negative sgRNAs (AUPRC: 20x=0.44, 50x=0.77, 100x=0.81, 200x=0.82). Further, at cell coverage of 20x, many sgRNAs fail to infect any cells, with high dropout rate (sgRNAs with less than 10 mapped reads in the expression sorting bins) of around 12%, and this value decreases to less than 1% as cell coverage greater than 50x (**Figure 3.2C**). Based on this analysis, we recommended at least 100x cell coverage for reproducible CRISPR screen design, but it should also be noted that even higher cell coverage may be needed for different CRISPR screen methods or for screening CREs for different genes or different types of cells.

Next, we tested the effect of varying sequencing depth to CRISPR screen accuracy by bootstrap sampling mapped sgRNA sequence reads, with simulated sequencing depth ranging from x5 to x1,000 (**Methods 3.2.4**). We first applied the sampling analysis to the cell coverage varying HCR-FlowFISH experiments, and found that with 250x sequencing depth or higher, the accuracy of HCR-FlowFISH screens for *GATA1* CREs is limited by cell coverage, such that further increase in sequencing depth only marginally improves accuracy (**Figure 3.2D**). We repeated the analysis in five other CRISPR screens, including growth screens performed at *GATA1* and *MYC* loci, finding 250x sequencing depth is a

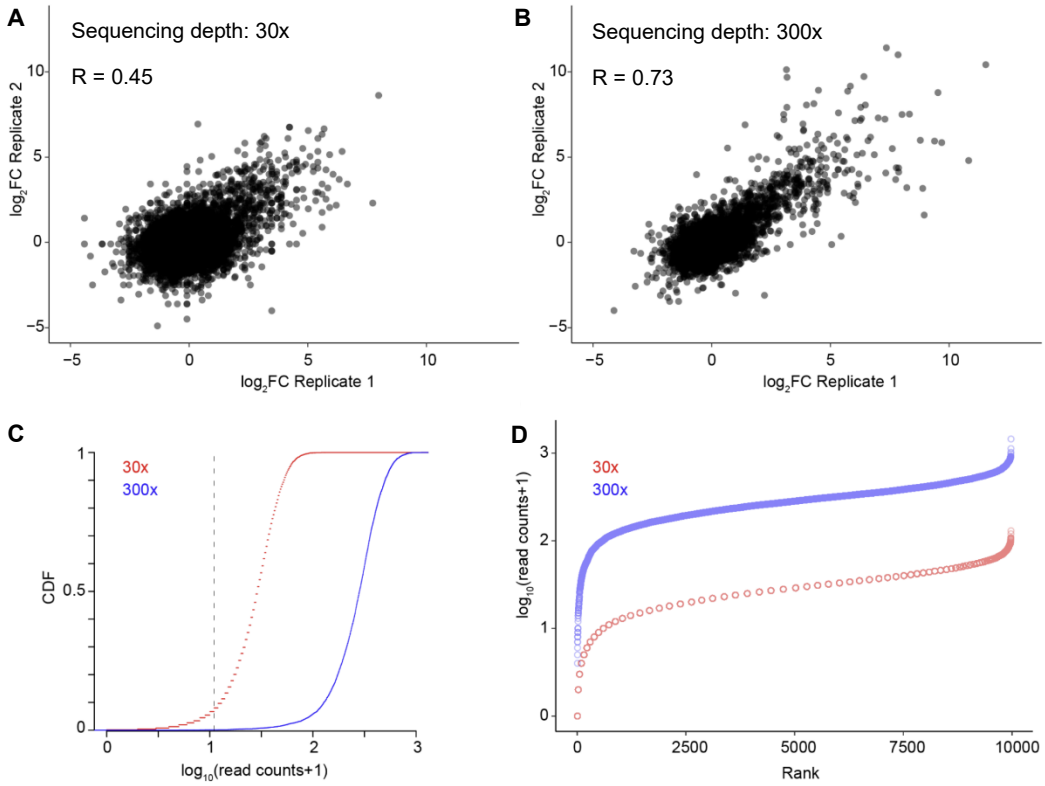

Figure 3.3. Representative bootstrap samples for low and high sequencing depths using K562 GATA1 locus CRISPRi growth screen

A) Biological replicate 1 log2FC vs Biological replicate 2 log2FC (Z-score) for 30x bootstrapped sequencing depth (9977 sgRNAs, R=0.45). **B)** Biological replicate 1 log2FC vs Biological replicate 2 log2FC for 300x (R=0.73). **C)** Empirical cumulative distribution function of log10 (sgRNA read counts) **D)** Dropout plot (rank of sgRNA read counts vs log10 (1+read counts) at 30x (red) and 300x (blue) bootstrapped sequencing depth.

reasonable minimum for CRE identification accuracy. We observed saturations of biological replicate correlation of guide effects (**Figure 3.2E**) and of guide dropout rate (**Figure 3.2F**) starting at 250x sequencing depth (biological replicate normalized log2FC R>0.9 and average dropout rate <2% for all screens). **Figure 3.3** shows representative bootstrap samples for low (30x) and high (300x) sequencing depth.

CRISPRi effects in the gene body are strand-specific

We also discovered that dCas9-KRAB can inhibit transcription when it targets the coding strand of a gene but not when targeting the template strand (**Methods 3.2.5**). We initially observed the strand-bias effect through one of the tiling GATA1 CRISPRi growth screens[132]. Intriguingly, among sgRNAs targeting the GATA1 gene body, those mapping to the positive strand of the reference genome had significantly higher perturbation effect sizes than those mapping to the negative strand. As a result, CRE calls made by CASA[103] from the tiling CRISPRi data predominantly mapped to the positive strand (**Figure 3.4A**). The positive and negative strands correspond to the coding and template strands for genes with positive transcriptional direction, and we suspected that the strand-bias effect is associated with the action of RNA polymerase. Subsequently, we investigated CRISPRi HCR-FlowFISH screens to confirm the consistency of our observation. We first looked into the FADS locus; it contains three homologous FADS genes, each of which is screened for CREs with tiling CRISPRi HCR-FlowFISH (**Figure 3.4B**). Consistent with our hypothesis, FADS1 (negative transcriptional direction) and FADS2 (positive transcriptional direction) each had higher CRISPRi perturbation effect for sgRNAs targeting the negative and positive strand of the corresponding gene bodies, implying that both growth screens and HCR-FlowFISH have coding-strand specific CRISPRi perturbation effect (**Figure 3.4C**). We verified that this strand-bias effect

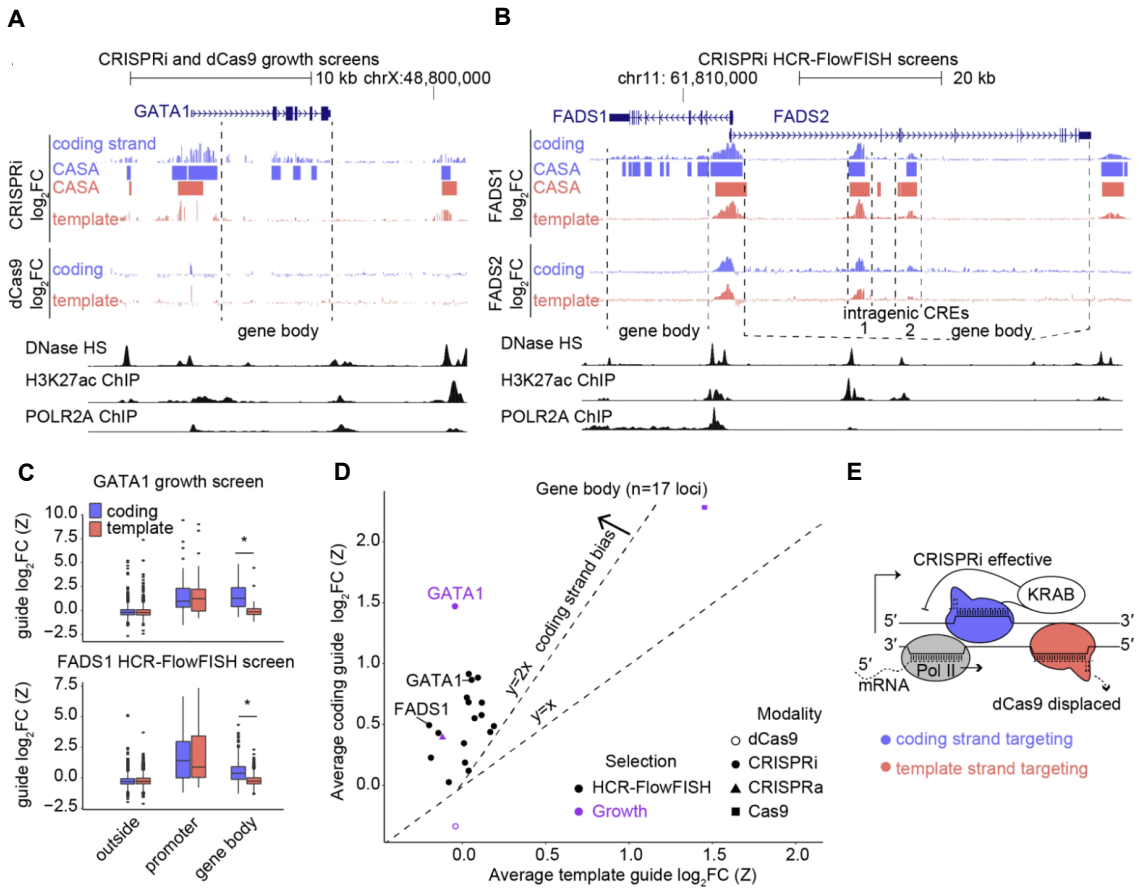

Figure 3.4. CRISPRi effects in the gene body are strand-specific

A) Strand-specific CRISPRi growth screen effects tiling GATA1. CRISPRi and dCas9 tracks show the average of two biological replicates, comparing Day 21 to plasmid (N=2,541 coding- and 2,263 template-strand targeting sgRNAs). **B)** Strand-specific CRISPRi HCR-FlowFish screen effects tiling FADS1 and FADS2. CRISPRi tracks show the average of two biological replicates, comparing high- versus low-expression bins for the target gene (N=4,609 and 4,942 sgRNAs per strand). **C)** Distributions of sgRNA effects (average of 2 screen replicates) in the gene body, and at the promoter (within 2 kb upstream of TSS), when sgRNAs are categorized by target strand, in the (top) GATA1 CRISPRi growth screen (n=2026, 1731, 34, 27, 100, and 77 sgRNAs from left to right boxes) and the (bottom) FADS1 HCR-FlowFish screen (n=3121, 3249, 90, 69, 520, and 702 sgRNAs). Boxes show the quartiles, with a line at the median, vertical lines extend to 1.5 times the interquartile range, and dots show outliers. Asterisk denotes significance with P<1e-15 by T-test. **D)** Strand-specificity across screens tiling 17 loci for sgRNAs targeting the gene body. Each point is the average effect of all sgRNAs from a screen targeting that region, averaged across two screen biological replicates, with color indicating the phenotypic readout, and shape indicating the type of CRISPR perturbation. **E)** Proposed model of gene body strand bias, wherein dCas9 binding could be reduced on the template strand due to competition with Pol-II-mediated transcription, rendering KRAB ineffective. In contrast, when targeting the coding strand, KRAB can be effective.

is consistent across all the other genes that had been screened with CRISPRi HCR-FlowFISH. For all the 17 genes, we observed a higher average effect size (log2FC Z-score) for coding-strand targeting guides than for template-strand targeting guides by at least two-fold (**Figure 3.4D**). Interestingly, this effect is specific to CRISPRi (dCas9-KRAB); dCas9 GATA1 growth screen exhibited an opposite effect, with a higher template strand effect size (although weak) than the coding strand in the gene body. We hypothesized that gene body targeting dCas9-KRAB and dCas9 inhibit transcription through two distinct mechanisms of actions. dCas9 blocks RNA polymerase (which 'walks' on the template strand) through steric hinderance, leading to a higher perturbation effect when dCas9 binds to the template strand. On the other hand, dCas9-KRAB induces a heterochromatin state in the gene body and reduces the efficiency of the polymerase (**Figure 3.4E**). The effectiveness of this inhibitory mechanism by dCas9-KRAB is higher when it targets the coding strand due to a lower collision rate with polymerase, resulting in a higher induction of repressive heterochromatin state of the gene body by dCas9-KRAB. This observation has an important implication for identifying intronic enhancers with CRISPR. For example, an intronic DHS with no regulatory interaction with the gene (containing the DHS in its gene body) can be mistakenly identified as a CRE hit for the gene if the gene-body coding-strand perturbation effect is not considered. To avoid such false-positive intronic CRE calls, intronic CRE hits should be called only for putative CREs with high CRISPRi perturbation effect sizes in both the template and coding strands.

3.3.2 Modeling regulatory dynamics of cell-state transition through noncoding CRISPR screen

Nuclear concentration of TFs determines the genome-wide gene expression profile, and thus precise regulation of TF expression levels is necessary for proper cell-state

transitions and normal development of multicellular organisms. Using the cell-state transition from human embryonic stem cells (ESC) to definitive endoderm (DE) as a model system, we sought to elucidate regulatory dynamics involved in cell-state transitions. Based on multiple lines of evidence, we identified the core TF regulators of the ESC (OCT4, SOX2, NANOG) and DE (EOMES, SOX17, GATA6) cell-states. Dissecting the regulatory process of the DE TF expression has significant medical implication, as mutations in TFs that regulate the DE differentiation are known to disrupt the normal development of multiple organs and cause diseases (e.g., GATA6 mutation can disrupt normal pancreatic development and cause diabetes[133]). We designed CRISPRi screen targeting putative CREs of the core TFs. We delivered dCas9-KRAB to each of the putative CREs, and quantified their roles as drivers of ESC to DE differentiation, by measuring the change in DE differentiation rate for each CRE perturbations. One of the key results of the screen is that perturbing the enhancers of DE TF weakly affects the steady-state cell-transition rate. That is, with sufficient amount of time of DE state induction (72 hours; DE-72h), most ESCs successfully differentiate to DE. However, if we measure the cell-transition rate at earlier time points (36 hours; DE-h36), ESCs with DE-TF enhancer perturbations show a decrease in the rate of cell differentiation. Hence, enhancer perturbation only leads to time-delay in cell-state transition, suggesting that the cell-state transition can be modeled as a bistable switch, where the ESC and DE states represent two stable points. We developed a nonlinear kinetic equation that closely mimics the cell-transition process revealed through the CRISPR screen. We also confirmed using Gillespie stochastic simulation, which is considered to be more biologically realistic, that the mechanism of cell-state transition depicted by the kinetic equation is biologically feasible. Together, we present a generalizable strategy that combines CRISPR functional characterization and mathematical modeling to uncover mechanisms of cell-state transition. In this section, I will briefly describe the experiments performed by Huangfu lab, and primarily focus on

computational modeling. For more comprehensive description, please refer to the original publication of this study[8].

CRISPRi screen reveals CREs that regulate the differentiation dynamics of ESC to DE

We CRISPRi-screened putative CREs of TFs that regulate the differentiation of ESC to DE. Based on a previous genetic screen[113], we have selected EOMES, MIXL1, GATA6, and SOX17 as the core DE TFs that drive the cell-state transition (**Figure 3.5A**). We further confirmed using RNA-seq and ATAC-seq that the activation of these TFs correlates with the transition of ESC to DE[8]. To quantify the impact of CRE perturbation to DE differentiation, we generated an ESC line with stable dCas-KRAB expression and GFP-tagged SOX17 gene, where SOX17-GFP expression was used as the cell marker for DE cell-state. We designed sgRNAs that target putative CREs of the four DE TFs, three ESC TFs (NANOG, OCT4, SOX2), and three signaling TFs (SMAD2, SMDA4, JUN) that are important for DE differentiation. We identified distal 394 ATAC-seq peaks, active in ESC or DE, within 4 million bps surrounding the TFs, and designed a total of 11,050 sgRNAs targeting these putative CREs. The sgRNAs were delivered to the ESCs with lentivirus (35 million ESCs; MOI~0.3), and induced the infected ESCs to differentiate into DE with Activin-A (TFB-β family ligand) for total 72 hours. At midpoint of the transition (36 hours), the cells were sorted with fluorescence-activated cell sorting (FACS) by SOX17-GFP expression, we used the cells with bottom 20% (SOX17-GFP^{+}) and top 20% (SOX17-GFP^{-}) for sgRNA enrichment analysis. sgRNAs that integrated into the two groups of cells were PCR-amplified and sequenced at 1000x depth, and the sgRNA sequencing reads were mapped to the sgRNA library. Then, enrichment of each guide in the SOX17-GFP^{-} cells relative to the SOX17-GFP^{+} cells was computed using guide quantities in each cell

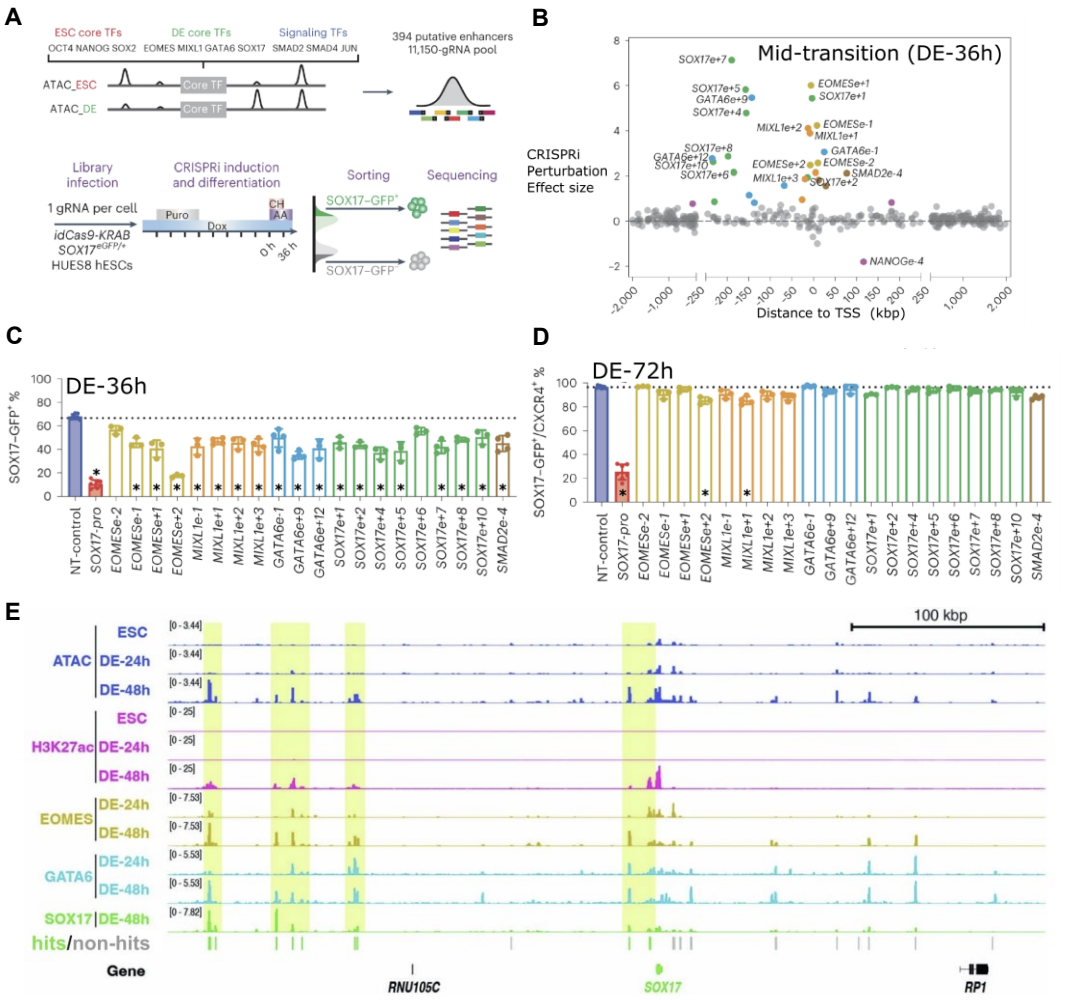

Figure 3.5. CRISPRi screen reveals CREs that regulate the differentiation dynamics of ESC to DE

A) The design of core enhancer CRISPRi perturbation screening **B)** The scatter plot of the CRISPRi perturbation effect size (Z-score) of reducing ESC-DE transition rate at the mid-transition point (DE-36h). Each dot represents putative enhancers targeted in the screen. TSS: transcription start site. **C)** Individual validations of the guide RNAs that target the top 20 enhancer hits in **B**. pro: promoter. Error bars indicate mean ± s.d. Statistical analysis was performed by two-tailed unpaired multiple comparison test with Dunnett correction. Asterisk (*) indicates p-value<10^{-4}. **D)** Same as **C** but measured at DE-72h. **E)** genome-browser view of the SOX17 locus showing ATAC-seq, H3K27ac ChIP-seq, and DE-TF (SOX17, GATA6, EOMES) ChIP-seq signals along with the enhancer hits identified through our CRISPR screen.

group (log_2FC), and computed the z-score of log_2FC for guide effect size. Lastly, we quantify the perturbation effect size of each CRE by calculating the average z-score of sgRNAs that target each CREs, resulting in 29 enhancer hits that regulate DE

differentiation (z-score: 0.75-7.14; **Figure 3.5B**). Most of the enhancer hits were in the DE gene loci, ranging from 4 to 9 enhancer per gene. For validation, we further individually tested the sgRNA perturbations targeting the top 20 enhancer hits and confirmed that the sgRNAs lead to downregulation of the cognate gene expression and also of the downstream SOX17-GFP signal (**Figure 3.5C**). Interestingly, the same perturbations had insignificant effect at DE-72h, implying that detection of the enhancers are temporally sensitive (**Figure 3.5D**).

These temporal enhancers dynamically activate after the DE-36h cell-transition point. For example, the SOX17 enhancers we identified (highlighted in **Figure 3.5E**) lack active enhancer marks prior to the DE-36h transition point (ESC and DE-24h). However, these enhancers gain strong ATAC-seq and H3K27ac ChIP-seq signals after the DE-36h transition point (DE-48h), and further bind with DE-TFs such as GATA6, EOMES and SOX17 itself. Similar observations were made in other DE-TF loci. These observations suggest that the DE-TFs drive ESC-DE cell-state transition through a cooperative autoregulatory network, motivating our computational model described below.

Modeling cell-state transition as a bistable switch recapitulates the temporal dynamics exhibited in the ESC-DE CRISPR screen

Based on the results of the CRISPRi screen, we modeled the dynamics of the cell-state transition as a bistable switch, having the ESC and DE states as the two stable points. In the original publication of the study, we constructed a continuous rate equation to model the system for mathematical analysis of the system. Since continuous rate equations are known to be more accurate at the high concentration limit, we also modeled the system as a discrete dynamical system, simulated with the Gillespie algorithm, and confirmed that

the mathematical characterization of the cell-state transition depicted by the continuous model is also reflected in the analogous discrete dynamical system. In this dissertation, I will take the reverse approach. I will first model the cell-transition with the discrete stochastic model, with explicit description of individual reactions involved in transcription (e.g., TF binding). I will use this model to run Gillespie simulation and recapitulate the CRISPR screen result exhibiting the temporal sensitivity. Then, I will describe a continuous kinetic equation having a similar mathematical structure with the discrete model, and use the kinetic equation for mathematical characterization of the ESC-DE cell-state transition.

I will first construct a discrete stochastic system that shows regulatory dynamics that recapitulates the observation in the ESC-DE model system, using the Gillespie algorithm. The Gillespie algorithm is a simulation method that realistically captures the stochastic fluctuation of molecules undergoing chemical reactions, especially for modeling reactions occurring in systems composed of small number of molecules[134]. To simulate a dynamical system with N reactants ($S_1, \dots, S_N$), the Gillespie algorithm requires specification of the initial number of each reactant ($X_1(t=0), \dots, X_N(t=0)$), all possible reactions that can occur in the system ($R_1 \dots, R_M$) and their effects to X_i, and propensity of occurrence for each reactions ($\alpha_1(X_i, \dots, X_N), \dots, \alpha_M(X_i, \dots, X_N)$). At each simulation step, the algorithm randomly selects reaction R_i with probability of $\frac{\alpha_i}{\sum \alpha_i}$, updates X_i based on the specified effect of R_i, and increments the time t to $t+\tau$, where τ is sampled from the exponential distribution with mean $\frac{1}{\sum \alpha_i}$. To build the simplest possible discrete stochastic system that recapitulates the temporal effect observed in the CRISPR screen, let us first consider a system of six types of transcription factors as illustrated in **Figure 3.6A**, where the three ESC TFs (red; SOX2, NANOG, OCT4) cooperatively regulate transcription of

one another through enhancer binding. Similarly, the DE TFs (green; SOX17, EOMES, GATA4) cooperatively regulate one another. The ESC TFs synchronously inactivates, and

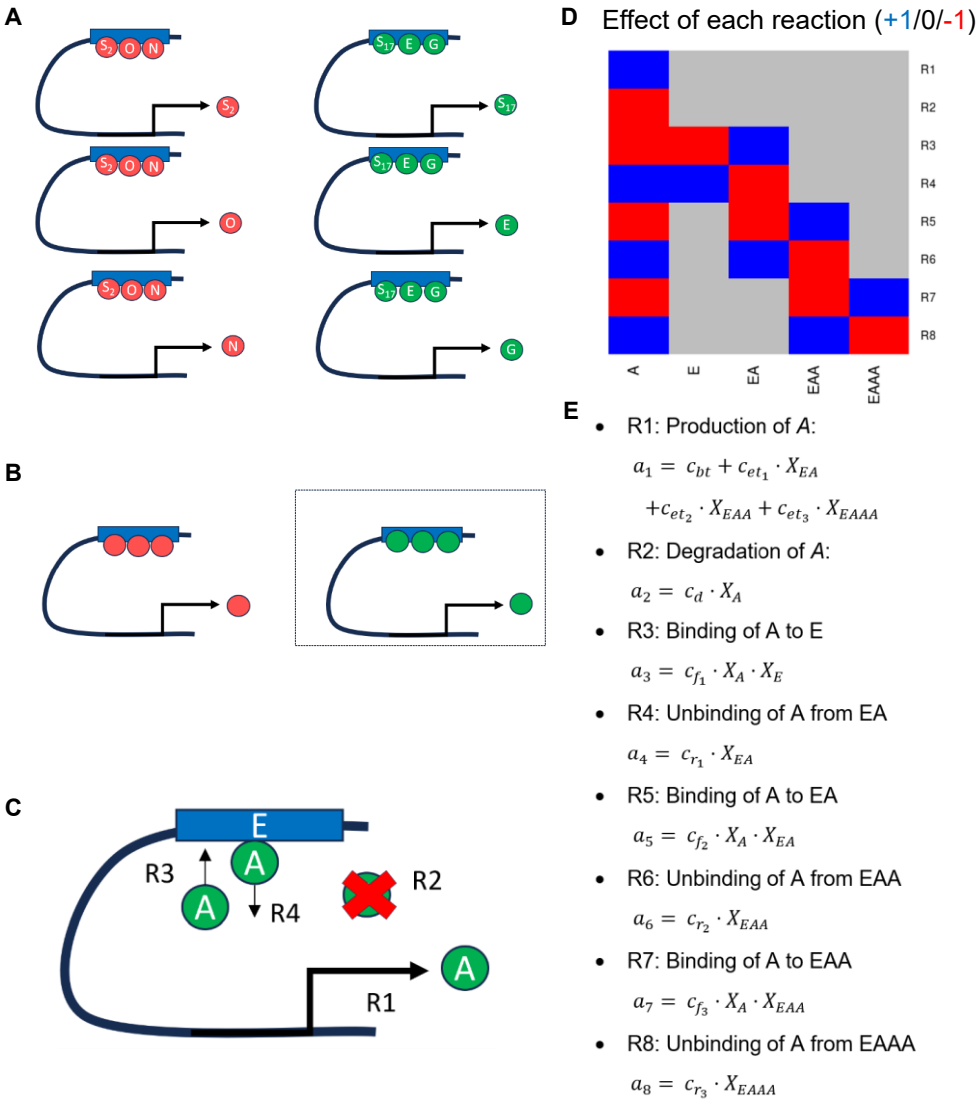

Figure 3.6. Discrete stochastic model of cell-state transition for Gillespie simulation

A) Cooperative autoregulation among ESC-TFs (red; S_2: SOX2, N: NANOG, O: OCT4) and among DE-TFs (green; S_{17}: SOX17, E: EOMES, G: GATA4). **B)** simplified model of the network under synchronous inactivation of ESC-TFs and activation of DE-TFs. The boxed portion indicates the DE-TF network. **C)** Components of the discrete stochastic model. A and E each represent TF and enhancer. R1-R4 denote some of the possible reactions in this model (R1: TF production, R2: TF degradation, R3: TF binding to enhancer, R4: TF unbinding from enhancer) **D)** Effect of each reaction to the system. For example, R1 increases the number of A by 1. **E)** propensity of occurrence (α) for each reaction. For example, the propensity of R2 (TF degradation) is proportional to the number of the TF in the system.

the DE TFs also synchronously activate through ESC-DE transition, which allows us to further simplify the dynamical system as illustrated in **Figure 3.6B** (single ESC gene, single DE gene; cooperative binding of indistinguishable TFs). Since our CRISPR screen focuses on the activation of DE cell-state but not on the inactivation of the ESC cell-state, we can focus on the boxed portion of **Figure 3.6B**, and construct a discrete dynamical system as depicted in **Figure 3.6C**. This system contains a total of 8 reactions (**Figure 3.6DE**), including production (R1; encompassing both the transcription and translation processes) and degradation (R2) of the TF (labeled as A) and the sequential binding and unbinding (R 3-8) of the TF to the enhancer (E). The production rate of the TF (reaction R1) is the sum of the basal transcription rate (c_{bt}) and the TF-enhanced transcription rates ($c_{et1} \cdot X_{EA} + c_{et2} \cdot X_{EAA} + c_{et3} \cdot X_{EAAA}$), and the strong cooperativity of enhancer activity is reflected in the fact that we chose $c_{et3} \gg c_{bt}, c_{et1}, c_{et2}$. Fifty independent simulations, each representing a single cell, were run. The initial enhancer state at $t = 0$ is given by $(X_E, X_{EA}, X_{EAA}, X_{EAAA}) = (1,0,0,0)$, and the initial number of TFs, X_A, was sampled from a uniform distribution. The stimulus (Activin-A for DE induction) was modeled by increasing in the initial rate of A binding to E as follows (sigmoidal activation):

$$c_{f1} \leftarrow c_{f1} \cdot s(t), \;\; where\; s(t) = \frac{s_o}{1 + e^{-\frac{1}{400}(t - \frac{t_{max} - t_{min}}{3})}}$$

Unless noted otherwise, we used the following rate constants for the propensity functions:

$$\left(c_{bt}, c_{et1}, c_{et2}, c_{et3}, c_{f1}, c_{r1}, c_{f2}, c_{r2}, c_{f3}, c_{r3}, c_d\right)$$
$$= (0.04, 0.04, 0.04, 2.4, 0.1, 15, 0.1, 15, 0.1, 15, 0.003)$$

The Gillespie simulation exhibits temporal dynamics analogous to the normal or CRISPRi-perturbed ESC-DE cell-state transition observed in our experiments. The simulated cells behave as a bistable switch, where the cells make transition to the high

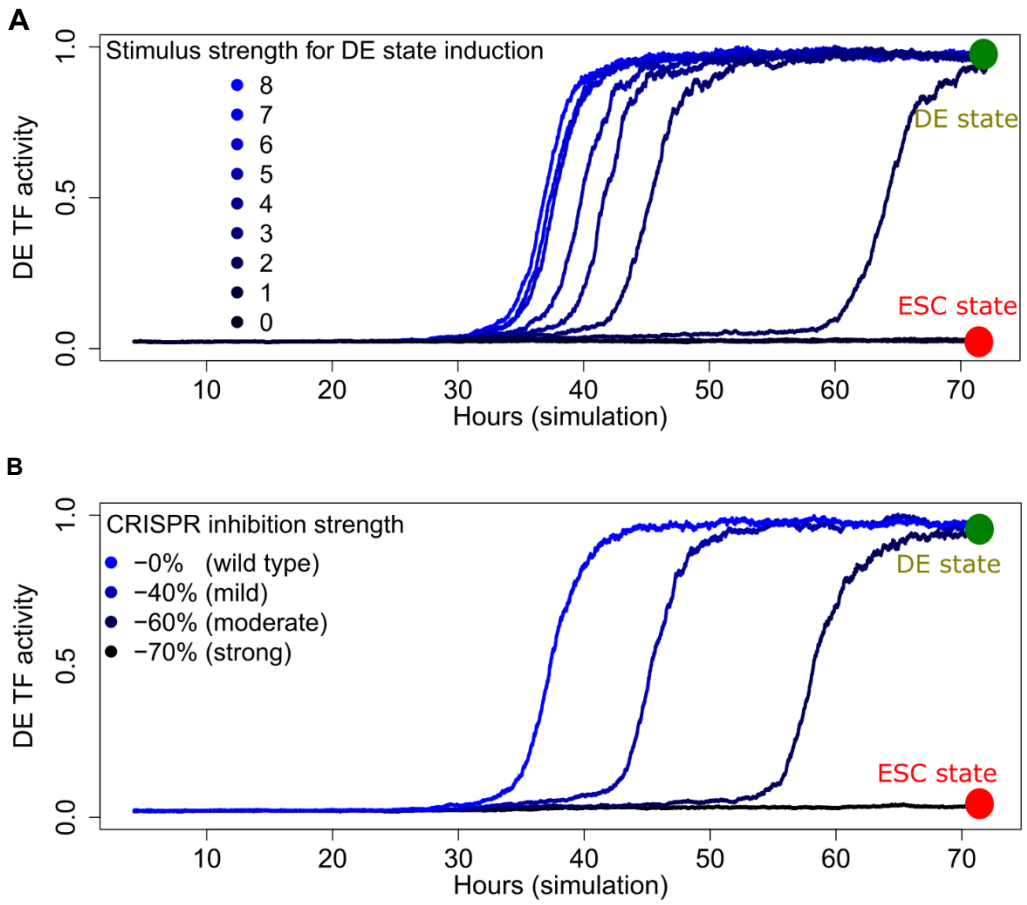

Figure 3.7. Gillespie simulation of the discrete model recapitulates the CRISPR-perturbed ESC-DE transition dynamics

(A,B) median #TF across 50 independent runs of Gillespie simulation **A)** with varying stimulus strength (S_o) and **B)** with varying degree of percent decrease in enhancer strength (c_{f1})

TF-expressing state (corresponding to the DE state) when the stimulus strength passes a certain threshold ($S_0 \sim 1.5$; **Figure 3.7A**). Below the threshold, the cells remain in the low-TF expressing state (corresponding to the ESC state) even through prolonged stimulus input. Above the threshold, further increase in the stimulus strength increases in the speed of cell-state transition, but the final TF level in the DE state remains constant. The simulation also shows the effect of enhancer perturbation, which we modeled as a reduction in TF binding rate to the enhancer (c_{f1}) (**Figure 3.7B**). The simulation shows that, with a perturbation strength higher than a certain threshold ($c_{f1} < 0.04$;

corresponding to more than 60% reduction), cells fail to make transition even with a strong stimulus ($S_0 = 7$). We have shown that such high level of perturbation may be achieved through CRISPR-Cas9-mediated deletion of multiple enhancers. Below the perturbation threshold ($c_{f1} \geq 0.04$; i.e. less than 60% reduction), most cells successfully make cell-state transitions, with transition speed highly variable with the perturbation strength. The cell transition speed is slower with higher perturbation strength, and this is analogous to the time-delay effect in ESC-DE differentiation that we observed with CRISPRi perturbations. At mid-point of the simulation, the transition rate is highly variable by the perturbation strength, but at the final-point, the transition rate is indistinguishable in both the experiment and the simulation.

For mathematical analysis of this temporal behavior, we can convert the discrete stochastic model into a continuous deterministic rate equation, which requires the assumption of rapid binding equilibrium between TF and enhancer. Since we set $c_{f1} = c_{f2} = c_{f3}$ and $c_{r1} = c_{r2} = c_{r3}$ for the simulations, we can write the following equations based on the binding equilibriums:

$$c_f \cdot X_A \cdot X_E = c_r \cdot X_{EA}$$

$$c_f \cdot X_A \cdot X_{EA} = c_r \cdot X_{EAA}$$

$$c_f \cdot X_A \cdot X_{EAA} = c_r \cdot X_{EAAA},$$

Which leads to

$$\frac{X_{EAAA}}{X_{EAAA} + X_E} = \frac{X_A^3}{\left(\frac{c_r}{c_f}\right)^3 + X_A^3}$$

Since $c_{et3} \gg c_{et1}, c_{et2}$ (i.e. high active transcription rate at full TF binding), we can define $P_{active} \coloneqq \frac{X_A^3}{\left(\frac{c_r}{c_f}\right)^3 + X_A^3}$ to denote the fraction of active enhancers. We can generalize this expression to arbitrary level of cooperativity (n) and write:

$$P_{active} = \frac{X_A^n}{\left(\frac{c_r}{c_f}\right)^n + X_A^n}$$

In the discrete model, we modeled X_A is an integer. For the continuous model, we replace X_A with a continuous variable ψ that represents the degree of TF activation, which gives us:

$$P_{active} = \frac{\psi^n}{\left(\frac{c_r}{c_f}\right)^n + \psi^n}$$

Further, in the discrete model, we modeled enhancer perturbation or stimulus impact by scaling the TF binding rate c_f. Using the scaling factor k,

$$P_{active} = \frac{\psi^n}{\left(\frac{c_r}{k \cdot c_f}\right)^n + \psi^n}$$

$$= \frac{k^n \psi^n}{\left(\frac{c_r}{c_f}\right)^n + k^n \psi^n}$$

Replacing $\left(\frac{c_r}{c_f}\right)^n$ with b and k^n with c:

$$P_{active} = \frac{c \cdot \psi^n}{b + c \cdot \psi^n}$$

, where b encodes TF-enhancer interaction, and c models enhancer strength tunable with CRISPRi.

In the discrete model, the number of TF is affected with only R1 (TF production) and R2 (TF degradation), R1 is composed of active and basal transcription, and the propensity of R2 scales with TF level. Hence, we can write the following rate equation that is mathematically analogous to the discrete model:

$$\frac{d\psi}{dt} = -r \cdot \psi + e_1 \cdot P_{active} + e_0 \cdot (1 - P_{active})$$

$$= -r \cdot \psi + e_1 \cdot \frac{c \cdot \psi^n}{b + c \cdot \psi^n} + e_0 \cdot \left(1 - \frac{c \cdot \psi^n}{b + c \cdot \psi^n}\right),$$

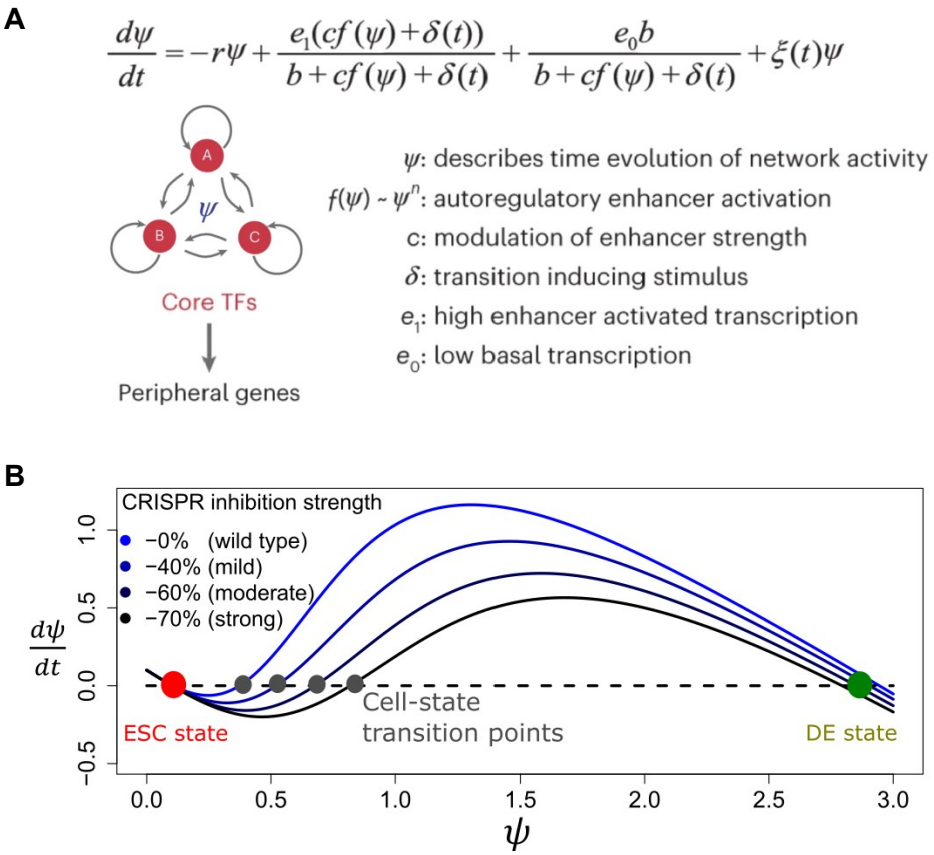

Figure 3.8. The continuous kinetic model explains the observed time-delay in ESC-DE transition.

A) Continuous kinetic equation for modeling cell-state transition. **B)** Phase portrait generated by plotting ψ vs $\frac{d\psi}{dt}$ using the kinetic equation. Each line drawn with varying enhancer strength (c) with the specified percent reduction.

, where r, e_1, e_0 are each the TF degradation rate, active transcription rate, and basal transcription rate (**Figure 3.8A**). Plotting $\frac{d\psi}{dt}(\psi)$ with $(b, c, e_0, e_1, r, n) = (0.5,1,0.1,3,1,3)$ gives the phase portrait in **Figure 3.8B**. The phase portrait shows three fixed points (ψ: $\frac{d\psi}{dt}(\psi) = 0$), with two stable fixed points ψ_{ESC} and ψ_{DE} and one unstable fixed point (ψ_u). When $\psi < \psi_u$, ψ decreases towards the first stable point ψ_{ESC}, and when $\psi > \psi_u$, ψ increases towards ψ_{DE}. The phase portrait shows that the modulation of enhancer strength (c) weakly affects ψ_{ESC} and ψ_{DE}, analogous to the observation in the CRISPR screen where CRISPRi perturbation weakly affected the SOX17 expression level at DE-72h. On the other hand, decreasing enhancer strength with CRISPRi has significantly increases the value of the unstable point ψ_u, requiring much higher activation ψ. This indicates that enhancer perturbation with CRISPRi can delay cell-state transition without dramatically affecting the initial (ψ_{ESC}) and the final cell-states (ψ_{DE}).

3.4 Discussion

Targeted epigenetic perturbation with CRISPRi allows direct functional characterization of gene regulatory elements at the levels of both gene expression and phenotype. I collaborated with multiple ENCODE labs to set up the largest database of CRISPR screens to date. Using the database, we demonstrated that pooled noncoding CRISPR screens generate highly reproducible perturbation signals that are consistent across diverse modalities of CRISPR screens (**Figure 3.1**). Through computational analysis of the screens, we generated an experimental guideline for performing efficient and reproducible noncoding CRISPR screens (**Figure 3.2**). Further, we discovered the strand-bias effect of perturbing gene-bodies with CRISPRi, which has an important implication for detecting intronic enhancers with CRISPR screens (**Figure 3.4**). In

collaboration with Dr. Huangfu and her team, we designed a CRISPR screen to identify enhancers that drive differentiation of embryonic stem cells (ESC) to definitive endoderm (DE), by targeting ATAC-seq peaks near the core TFs of ESC and DE (**Figure 3.5**). Interestingly, perturbing DE-TF enhancers with CRISPRi only affected the cell-state transition speed from ESC to DE but not the final cell-state, exhibiting a high SOX17 expression level post-transition. These enhancers dynamically activate through DE-TF binding, suggesting that cooperative autoregulation among DE-TFs drive the ESC-DE transition. Simulation modeling this behavior closely matches experimental observations (**Figure 3.6, 3.7 3.8**). Together, this chapter detailed how CRISPR can be utilized to characterize enhancer functions, which will be valuable for elucidating causal mechanisms of diseases associated with regulatory variants.

Chapter 4

Modeling Human and Mouse Enhancer Sequences with Machine Learning

Enhancers are clusters of degenerate binding sites for transcription factors (TF), and deciphering the sequence structure of enhancers is a computationally dauting task yet indispensable for elucidating biological functions and molecular evolution of enhancers[4]. Enhancers are known to evolve rapidly, with a limited number of identifiable ortholog between human and mouse with conserved regulatory activities. The sequence and functional flexibilities of enhancers are the likely causes of high enhancer turnover rates, but the observed turnover rate may also be inflated by limitations in current computational methods used for mapping orthologous enhancers. To comprehensively profile enhancer evolution and identify causes of the apparent lack of enhancer conservation, we obtained 45 pairs of orthologous human and mouse cell/tissues with matched gene regulatory profiles from more than 1,000 ENCODE DNase-seq experiments across diverse human and mouse biosamples. The similarities in regulatory profiles between human and mouse samples were quantified by comparing their enhancer sequence models derived using gkm-SVM, a sequence-based machine learning model. This strategy of matching human and mouse cell/tissues was motivated by past observations that TF functions are well-conserved across animals and that enhancers are bound by TFs with cell-specific expressions.

[4] Most of the contents in this chapter (Figure 4.4-4.9) are adapted from the original gkm-align manuscript[10].

In this chapter, I will first provide a brief overview of gkm-SVM. I will provide its mathematical foundation and introduce gapped-kmers, a sequence feature utilized by both gkm-SVM and the novel *gkm-align* genome alignment algorithm (**chapter 7**). Then, I will show how enhancer sequence features can be extracted from enhancer-trained gkm-SVM. These features are encoded in kmer weights, where kmers with high weights represent enhancer TF binding motifs. For simplicity, I will call these kmers as enhancer *regulatory vocabularies*. Using simulated synthetic enhancers, I will demonstrate that gkm-SVM regulatory vocabularies encode TF binding motifs (or more generally, prevalent DNA sequence patterns in the training set) and show how they can be used to compare enhancer models obtained from different cell-types. Lastly, I will use this principle to compare diverse human and mouse cell/tissues by their enhancer and promoter sequence models, and demonstrate that enhancer regulatory vocabularies are well-conserved between human and mouse in a cell-specific manner[10]. The 45 cell/tissue pairs derived in this chapter will be used to comprehensively quantify enhancer evolution (**Chapter 5**) and to evaluate the novel *gkm-align* algorithm against conventional genome-alignment methods (**Chapter 6**).

4.1 Introduction to enhancer sequence modeling with gkm-SVM

Gkm-SVM is a SVM (support vector machine) based machine learning method that computes an optimal hyperplane that maximally separates positive (e.g., brain enhancers) and negative training sequences (e.g., random genomic sequences) by their gapped-kmer compositions. The resulting sequence model encodes gapped-kmers that can be linearly combined to predict regulatory elements. In this section, I will introduce SVM and gapped-kmers, and combine the two to derive gkm-SVM. Mathematical details given in this section

will also be used in **chapter 7**. For more technical detail and its algorithmic variations, please refer to the original gkm-SVM paper[23,135].

4.1.1 Support Vector Machine (SVM)

I will first give a brief overview of SVM (support vector machine)[136]. Suppose we are given *N* number of *m*-dimensional vectors, x_i, each of which are labeled as either positive or negative. This information can be represented as $(x_1, y_1), \ldots, (x_N, y_N)$, where $y_i = 1$ if x_i is a positive or $y_i = -1$ if x_i is a negative vector. SVM involves finding w and b such that $w \cdot x_i - \rho \geq 1$ if $y_i = 1$ and $w \cdot x_i - \rho \leq -1$ if $y_i = -1$ (i.e., $y_i(w \cdot x_i - \rho) \geq 1$). Geometrically, this problem is equivalent to finding a hyperplane that separates the positive and negative vectors with a margin. We can define a hyperplane as a set of all x_i such that $w \cdot x_i - \rho = 0$, where w is a vector perpendicular to the hyperplane and ρ is a scalar that translates the hyperplane along the direction of w by $\frac{\rho}{\|w\|}$. Let $\widehat{H} = \{x_i : w \cdot x_i - \rho = 0\}$ be a hyperplane that places the positive vectors on one side and the negative vectors on the other side. The goal of SVM is to find the optimal hyperplane $\widehat{H}$ that robustly separates the positive and negative vectors with maximal margin. Here, we define margin as the distance between the following two hyperplanes, $H_{(+)} = \{x_i : w \cdot x_i - \rho = 1\}$ and $H_{(-)} = \{x_i : w \cdot x_i - \rho = -1\}$, which are parallel to $\widehat{H}$ and symmetrically positioned around $\widehat{H}$. With a hard margin, the positive vectors will have $w \cdot x_i - \rho \geq 1$ and the negative vectors will have $w \cdot x_i - \rho \leq -1$, geometrically placing the positive and negative vectors on the two opposing sides of $H_{(+)}$ and $H_{(-)}$. Denoting the margin as c, the orthogonal distance between $H_{(+)}$ and $H_{(-)}$, is computed with $x_{(+)} - x_{(-)} = c\frac{w}{\|w\|}$, where $x_{(+)}$ and $x_{(-)}$ are each vectors in $H_{(+)}$ and $H_{(-)}$. Since $w \cdot x_{(+)} = 1 + \rho$ and $w \cdot x_{(-)} = -1 + \rho$,

$$w \cdot \left(x_{(+)} - x_{(-)}\right) = c\frac{w \cdot w}{\|w\|}$$

$$\rightarrow 2 = c\frac{w \cdot w}{\|w\|}$$

$$\rightarrow c = \frac{2}{\|w\|}$$

Hence, maximizing the margin c is equivalent to minimizing $\|w\| = \sqrt{w \cdot w}$. Applying the constraint that the positive and negative vectors must be separable by the margin, we obtain the following optimization problem.

$$argmin_{w,\rho}\|w\|^2 \quad s.t. \;\; y_i(w \cdot x_i - \rho) \geq 1$$

But in practice, a hyperplane that perfectly separates enhancers and non-enhancers does not exist. The soft-margin strategy transforms the constraint into a penalty, and the optimization problem converts to:

$$argmin_{w,\rho}\, \lambda\|w\|^2 + \frac{1}{N}\sum_{i=1}^{N} \max\big(0, 1 - y_i(w \cdot x_i - \rho)\big), \qquad \lambda > 0$$

λ (inversely related to C in the LIBSVM package[137]) is a hyperparameter of SVM, and increasing λ (or decreasing C) leads to higher relaxation of the constraint. Solving the above optimization problem leads to the following solution:

$$\widehat{w} = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\hat{\rho} = \widehat{w} \cdot x_i - y_i$$

where x_is with $\alpha_i \neq 0$ are defined as ***support vectors***. Then, SVM prediction for an input vector x is be computed as:

$$\hat{y} = \widehat{w} \cdot x - \hat{\rho} = \sum_{i=1}^{N} \alpha_i y_i (x_i \cdot x) \; - \hat{\rho}$$

(Eq. 4.1)

4.1.2 Modeling enhancers using gapped-kmers

In this section, I provide a list of definitions and annotations related to gapped-kmers. The final goal of this section is to precisely define how two genomic sequences can be compared by their gapped-kmer compositions. The gapped-kmer based sequence similarity metric is at the core of both gkm-SVM and the novel *gkm-align* algorithm, which will be discussed in **chapter 6 and 7**.

Kmers and gapped-kmers are defined as follows:

- Definition: a kmer(k) is a length-k sequence of $\{A, C, G, T\}$
- Definition: a gapped-kmer(l,k) is a length-l sequence of $\{A, C, G, T, -\}$ with k number of elements from $\{A, C, G, T\}$.

In other words, a gapped-kmer is a kmer with a fixed number of gaps (-). Gaps are wildcard characters. Denoting a gapped-kmer(l,k) as s_g and a kmer(l) as s_k,

- Definition: s_g is contained ($\in$) in s_k if $\left(s_g[i] = s_k[i]\right)$ or $s_g[i]$ is a gap (-) for all $1 \leq i \leq l$. s_k is contained ($\in$) in sequence S if s_k is a subsequence of S.

In other words, a gapped-kmer is said to be contained in a kmer if they are equal at all base pairs except at positions where there is a gap in the gapped-kmer. Using this definition, we can count the number of a specific gapped-kmer contained in a sequence ($N_{s_g \in S}$) by counting the number kmers (s_k), contained in the sequence (S), that contain the gapped-kmer (s_g). That is:

$$N_{s_g \in S} = \sum_{s_k \in S} \mathbf{1}_{s_k}(s_g),$$

where

$$\mathbf{1}_{s_k}(s_g) = \begin{cases} 1 \; if \; s_g \in s_k \\ 0 \; if \; s_g \notin s_k \end{cases}$$

Finally, let us define $g(S)$, a vector that encodes the number of gapped-kmers(l,k) contained in S, for each of the $\binom{l}{k}4^k$ possible gapped-kmers. $g(S)$ is a vector of size $\binom{l}{k}4^k$, where the ith element of $g(S)$ encodes the number of the ith gapped-kmer in alphabetical order. That is,

$$g(S) \coloneqq \begin{bmatrix} N_{s_{g_1} \in S} \\ N_{s_{g_1} \in S} \\ \vdots \end{bmatrix}$$

I will call $g(S)$ a gapped-kmer vector of S. Lastly, we can compute the similarity between sequence $S_1$and S_2 by comparing their gapped-kmer vectors with cosine similarity:

$$G(S_1, S_2) = \frac{g(S_1) \cdot g(S_2)}{\|g(S_1)\| \cdot \|g(S_2)\|}$$

(Eq. 4.2)

4.1.3 gkm-SVM: support vector machine based on gapped-kmer sequence features

Recall from the previous section that making SVM prediction on an input vector x involves computing the inner products between x and the support vectors.

$$\hat{y} + \hat{\rho} = \sum_{i=1}^{N} \alpha_i y_i (x_i \cdot x)$$

In gkm-SVM, $x = \frac{g(S)}{\|g(S)\|}$, a normalized gapped-kmer vector of a sequence. Then, SVM prediction for an input sequence S is computed as

$$\hat{y}(S) + \hat{\rho} = \sum_{i=1}^{N} \alpha_i y_i \frac{g(S)}{\|g(S)\|} \cdot \frac{g(S_i)}{\|g(S_i)\|}$$

$$= \sum_{i=1}^{N} \alpha_i y_i G(S, S_i)$$

(Eq. 4.3)

Hence, making enhancer predictions with gkm-SVM involves computing pairwise sequence similarities $G(S_1, S_2)$, which can be computed without counting the gapped-kmers to derive $g(S)$ as I show below. First, note that gapped-kmers in sequence S can be counted by summing up the number of gapped-kmers contained in kmers that are contained in S:

$$g(S) = \sum_{s_k \in S} g(s_k)$$

(Eq. 4.4)

Denoting a length-l subsequence of S starting at the ith position as $S(i, l)$,

$$g(S) = \sum_{s_k \in S} g(s_k) = \sum_{i=1}^{|S|-l+1} g(S(i, l))$$

Then,

$$g(S_1) \cdot g(S_2) = \left(\sum_{i=1}^{|S_1|-l+1} g\big(S_1(i, l)\big) \right) \cdot \left(\sum_{i=1}^{|S_1|-l+1} g\big(S_2(i, l)\big) \right)$$

$$= \sum_{i=1}^{|S_1|-l+1} \sum_{j=1}^{|S_2|-l+1} g\big(S_1(i, l)\big) \cdot g\big(S_2(j, l)\big)$$

Recall that $S(i, l)$ is a length-k kmer and $g()$ counts the number of length-l gapped-kmers with k non-gap characters. There can only be one or no gapped-kmer contained in a kmer

of same length, and therefore $g(S_1(i,l))$ is a binary vector, and thus $g(S_1(i,l)) \cdot g(S_2(j,l))$ counts the total unique gapped-kmers commonly contained in the two subsequences. This can be computed by calculating the number of ways that non-wildcard characters (A,C,G,T) can be placed at positions along the kmers with matching characters. For example, if $s_{k1} = AAAAAAAAAA$ and $s_{k2} = AAATAGAAAA$, the two kmers are mismatching at position 4 and 6, and only the gapped-kmers that have gaps at position 4 and 6 can fit in both kmers. For counting gapped-kmer(l=10,k=6), there are 4 wildcard characters out of 10. Two of the wildcard characters must be placed at the two mismatched positions, and the number of ways the remaining two wildcard characters can be placed at the eight matched positions is $\binom{\#matched}{(\#wildcard)-(\#mismatch)} = \binom{\#matched}{(\#matched)-((\#wildcard)-(\#mismatch))} = \binom{l-(\#mismatch)}{l-(\#wildcard)}$, where $k = l - (\#wildcard)$. Denoting the number of base pair mismatches as m, we obtain:

$$g(s_{k1}) \cdot g(s_{k2}) = \begin{cases} \binom{l-m}{k}, l-k \geq m \\ 0 \quad , l-k < m \end{cases}$$

Then,

$$g(S_1) \cdot g(S_2) = \sum_{i=1}^{|S_1|-l+1} \sum_{j=1}^{|S_2|-l+1} g(S_1(i,l)) \cdot g(S_2(j,l))$$

$$= \sum_{i=1}^{|S_1|-l+1} \sum_{j=1}^{|S_2|-l+1} \begin{cases} \binom{l-m_{ij}}{k}, l-k \geq m_{ij} \\ 0 \quad , l-k < m_{ij} \end{cases}$$

(Eq. 4.5)

m_{ij} denotes the number of mismatched base pairs between $g(S_1(i,l)) \cdot g(S_2(j,l))$. This equation for computing the inner product of gapped-kmer vectors is then used to compute the cosine similarity (Eq. 4.2). Gkm-SVM[23] efficiently computes the mismatch profile using

a kmer tree structure. Gkm-align[10] (**chapter 7**) computes the mismatch profile using SIMD parallel computation, which is more efficient for its algorithmic structure.

4.1.4 gkm-SVM regulatory vocabulary

After training gkm-SVM on enhancer sequences, we can extract sequence patterns that are enriched in enhancers relative to the random genomic sequences used as negative training sets. The enhancer sequence patterns are extracted in the form of a list of weighted kmer sequences for all possible length-l kmers. Ignoring the constant $\hat{\rho}$ term, the kmer weights are computed by making SVM predictions on each kmers:

$$w_{s_k} \coloneqq \hat{y}(s_k) + \hat{\rho} = \sum_{i=1}^{N} \alpha_i y_i G(s_k, S_i), \qquad using\ eq.4.3$$

(Eq. 4.6)

Now I show that the weighted kmers contain all necessary and sufficient information to make SVM enhancer prediction and also make predictions for variant impacts[3].

$$\hat{y}(S) + \hat{\rho} = \sum_{i=1}^{N} \alpha_i y_i \frac{g(S) \cdot g(S_i)}{\|g(S)\|\|g(S_i)\|}, \qquad using\ eq.4.3$$

$$= \sum_{i=1}^{N} \alpha_i y_i \frac{\left(\sum_{s_k \in S} g(s_k)\right) \cdot g(S_i)}{\|g(S)\|\|g(S_i)\|}, \qquad using\ eq.4.4$$

$$= \sum_{s_k \in S} \sum_{i=1}^{N} \alpha_i y_i \frac{g(s_k) \cdot g(S_i)}{\|g(S)\|\|g(S_i)\|}$$

$$= \frac{\|g(s_k)\|}{\|g(S)\|} \sum_{s_k \in S} \sum_{i=1}^{N} \alpha_i y_i \frac{g(s_k) \cdot g(S_i)}{\|g(s_k)\|\|g(S_i)\|}$$

$$= \frac{\|g(s_k)\|}{\|g(S)\|} \sum_{s_k \in S} \sum_{i=1}^{N} \alpha_i y_i G(s_k, S)$$

$$= \frac{\|g(s_k)\|}{\|g(S)\|} \sum_{s_k \in S} w_{s_k} \ , \qquad using\ eq.4.6$$

$$= \frac{\sqrt{\binom{l}{k}}}{\sqrt{(|S| - l + 1)^2 \binom{l}{k}}} \sum_{s_k \in S} w_{s_k} \ , \qquad using\ eq.4.5$$

$$= \frac{1}{|S| - l + 1} \sum_{s_k \in S} w_{s_k}$$

(Eq. 4.7)

Since $\frac{1}{|S|-l+1}$ equals the number of length-l kmers contained in S, gkm-SVM prediction can equivalently be computed as the average weights of kmers contained in the input sequence S, which is considerably faster than computing $G(S, S_i)$ for every support vector S_i for each SVM predictions. This allows enhancer prediction and also prediction of regulatory variant impact with kmer weights, as done in the original delta-SVM paper[3]. Denoting S_f as a DNA sequence upon mutagenesis in the initial sequence S_i, the regulatory impact of the variant is computed as:

$$\Delta\hat{y}(S) = \hat{y}(S_f) - \hat{y}(S_i) \propto \left(\sum_{s_k \in S_f} w_{s_k} - \sum_{s_k \in S_i} w_{s_k} \right)$$

(Eq. 4.8)

For a single nucleotide substitution of S_1at the p^{th}character, only the kmers that overlap with the p^{th}character contribute to the differences in the kmer weight sums, and all the other flanking kmer weights cancel out.

A kmer weight threshold can be arbitrarily chosen to decide which kmers are significantly enriched in enhancers and call them significant. But it should also be noted that kmers have a spectrum predictive power, with kmers encoding consensus TF binding motif having the highest weights while kmers encoding degenerate TF binding motifs having relatively lower kmer weights. Therefore, even though I will use the term regulatory vocabulary for convenience, the full range of kmers need to be considered for a complete modeling of enhancer sequences. I refer to the full kmer-weight from gkm-SVM training as *regulatory vocabulary*. Later in this chapter, I will quantify the similarity of enhancer regulatory vocabulary between two cell/tissues by comparing their full kmer weight distributions, for example, with Pearson correlation.

4.2 Methods

4.2.1 Nucleotide entropy in simulated enhancers

Nucleotide frequency for each base pair positions is first computed for a given set of simulated enhancers with matching TFBS positions. For the ith sequence position, nucleotide entropy as computed as

$$S_i = -\sum_{n\in\{A,C,G,T\}} P_i(n)\ln\big(P_i(n)\big),$$

where $P_i(n)$ is the frequency that the ith nucleotide equals $n \in \{A, C, G, T\}$.

4.2.2 delta-SVM

Delta-SVM score quantifies the change in gkm-SVM prediction upon a change in sequence. Delta-SVM for a single nucleotide variation (SNV) is computed with Equation

4.8. The delta-SVM scores for all the three possible nucleotide changes are averaged to make delta-SVM binding site prediction, and I will call this mean delta-SVM score.

4.2.3 Annotating kmers by TF binding motifs

Each kmer was quantified for sequence relevance for each of the five TF-binding PWMs used for the simulation (SOX17, EOMES, GATA6, SMAD4, TCF4). The likelihood of observing each kmer under a given TF-binding PWM was computed, across the full length of each kmer, using uniform distribution for PWM positions not overlapping with a kmer nucleotide. I used the maximum kmer likelihood, among the PWM positions, to quantify sequence association of each kmer to each of the TFs.

4.2.4 Generating enhancer and promoter sets from DNase-seq data and gkm-SVM training

For all the ENCODE DNase-seq data used, DNase-seq filtered alignment bam files were downloaded from the ENCODE portal https://www.encodeproject.org/. For the primate fibroblast DNase-seq data, primate fibroblast DNase-seq raw fastq files (under GEO accession GSE129034) were downloaded from (Edsall 2019)[138] and mapped to chimpanzee (panTro6), gorilla (gorGor5), orangutan (ponAbe3), and rhesus (rheMac8) genomes using bowtie2[87] (-L 20). DHS peaks were called by running MACS2[88] on bams from combined replicates using default parameters but more stringent p-value (-p 1e-9). 300 bp peaks were generated by extending +/- 150bp from MACS2 summits.

Promoter sets were generated by selecting all peaks within 2kb of any annotated TSS. Remaining distal cell specific peaks were further filtered by removing all peaks called in more than 30% of ENCODE DHS samples (hg38ubiq30.bed and mm10ubiq30.bed). This definition of promoters and enhancers was used to train the promoter and enhancer gkm-

SVM models that are publicly available on the ENCODE portal (**Supplementary Table 2-3** for ENCODE accession IDs). All the supplementary tables are accessible through beerlab.org/gkmalign/. To define primate enhancers, coordinates in hg38ubiq30.bed were mapped to the respective primate genomes by LiftOver. For all the analysis in the gkm-align manuscript[10] (**chapter 4-5**), more stringent definition was used for defining promoter elements to filter out proximal DHSs that are accessible in less than 30% of DHS samples.

For gkm-SVM model generation, gkm-SVM models were trained on top 10,000 300bp enhancers or top 10,000 300bp promoter peaks (with highest DNase-seq MACS2 peak score)vs GC and repeat matched random sequence following (Beer Shigaki Huangfu 2020; Ghandi 2014; Lee 2015)[3,23,24] using default parameters (-l 11 -k 7).

To generate gkm-SVM kmer-weight vectors, A fasta file containing all 11-mers was generated using nrkmers.py from the lsgkm software package (v0.1.1) package[135]. Kmer weight vector for each 1,270 human and 153 mouse gkm-SVM enhancer models were generated using the lsgkm package's *gkmpredict* on the 11-mer fasta file using respective gkm-SVM models.

4.2.5 Visualization of epigenetic signals at DHSs

We used deepTools's computeMatrix (v3.5.1)[139] to compute average epigenetic signals across all DHS for each -1000 to 1000 base pair positions (resolution: 10) relative to DHS center using the following command:

computeMatrix reference-point -S example.bigWig -R example.bed -a 1000 -b 1000 --referencePoint center -bs 10 -o example.tab.gz

ENCODE ChIP-seq and DNase-seq bigwig files used for this analysis (**Figure 4.5.B**, **Figure 4.6**) are listed in **Supplementary Table 1**.

4.2.6 Interspecies cis-regulatory element prediction

Top 10,000 elements for each DHS class (enhancer, promoter) in human and mouse were used as *positive* set, and randomly sampled 300bp genomic loci with matched GC-content and repeat annotations were used as *negative set*. Gkm-SVM model trained in one species was used to make prediction for elements in each *positive* and *negative* set in the other species. AUROCs were computed from these interspecies prediction values.

4.2.7 Generation of 45 human-mouse cell/tissue pairs

For each human sample (N = 1,270; **Supplementary Table 2**), we identified its best matching mouse sample (N = 153; **Supplementary Table 3**) by finding mouse sample with the highest gkm-SVM kmer weight Pearson correlation. The best matching human sample for each mouse sample was identified similarly. The 45 human-mouse cell/tissue pairs were derived as those human and mouse samples that reciprocally mapped to each other by top matched kmer weights (**Supplementary table 4**). All the supplementary tables are accessible through beerlab.org/gkmalign/.

4.3 Extracting simulated regulatory vocabularies from synthetic enhancers

In this section, I will demonstrate how gkm-SVM extracts enhancer regulatory vocabularies. I will use gkm-SVM to extract simulated regulatory vocabularies from synthetic enhancers, which are generated by seeding TF motifs using position weight matrices (PWM). This exercise, together with past experimental validations[3,28,30] of gkm-SVM regulatory vocabularies, will justify the usage of gkm-SVM for quantifying regulatory conservation between human and mouse in the next section.

I simulated synthetic enhancers by seeding TFBS to random genomic sequences (**Figure 4.1A**). I chose five transcription factors active in definite endoderm[24] (SOX17, EOMES, GATA6, SMAD4, TCF4), and used publicly available information on the TFs' DNA binding preferences, encoded in PWMs[140]. For each of the five TFs, I used its PWM to sample its binding sequences (length 9-12 base pairs). I sampled 10 different types of enhancers by choosing three out of the five TFs and separating their DNA-binding sequence samples (PWM) by 5 bp-long spacers. 1,000 TFBS combinations were generated for each of the 10 enhancer types, and these sequences were seeded to the centers of 10,000 randomly sampled 300 base pair wide human genomic sequences (i.e. base pairs replaced at TF binding positions but not at spacers or the flanking sequences). This leads to 10,000 sequences of 300bp-long enhancers, with 10 different cis-regulatory modules seeded at their centers (**Figure 4.1A**). For comparison, I also built a simpler set

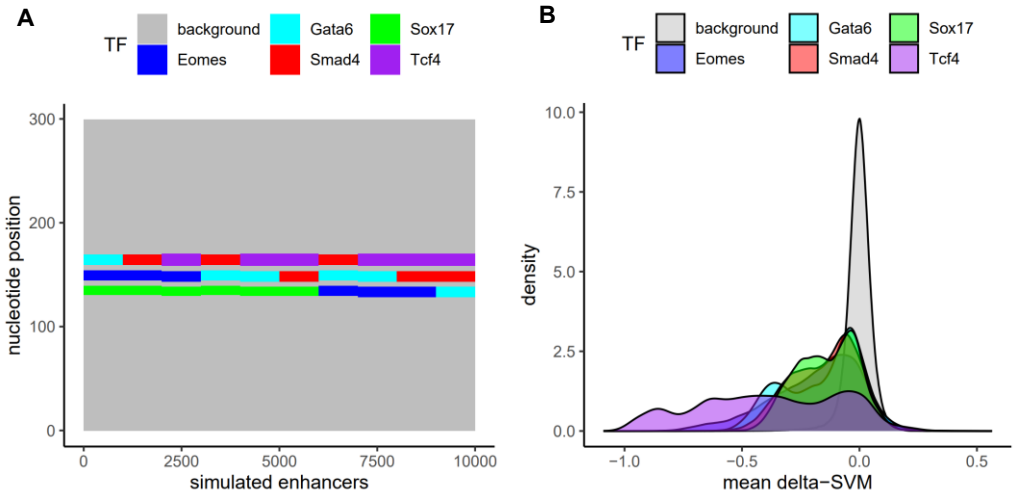

Figure 4.1. gkm-SVM regulatory vocabulary encodes sequence patterns enriched in the positive training sequence set.

A) Visualizing 10 classes of synthetic enhancers with various combinations of seeded PWM TF motifs. Total 10,000 enhancers (x-axis) of width 300bp (y-axis) were simulated. Each base pair position is colored by seeded PWM TF motif. **B)** Mean delta-SVM (averaged across all possible nucleotide variations) for all 10,000 x 300 positions in the synthetic enhancers, grouped by seeded PWM TF motif as visualized in **A**.

of enhancers by choosing two TFs out from SOX17, EOMES, GATA6, leading to three enhancer types with 3,333 sequences each (**Figure 4.3A**)

To demonstrate that gkm-SVM regulatory vocabularies recover the seeded TF binding patterns, I trained gkm-SVM on these 10,000 synthetic enhancers with diverse TFBS patterns (**Figure 4.1A**) against another set of 10,000 randomly sampled genomic sequences. Then, I extracted the regulatory vocabularies of the simulated enhancers (Eq. 4.7), and used them to make TF binding site prediction (Eq. 4.8) along the simulated enhancer sequences. I computed nucleotide entropy for each of the 10 enhancer types,

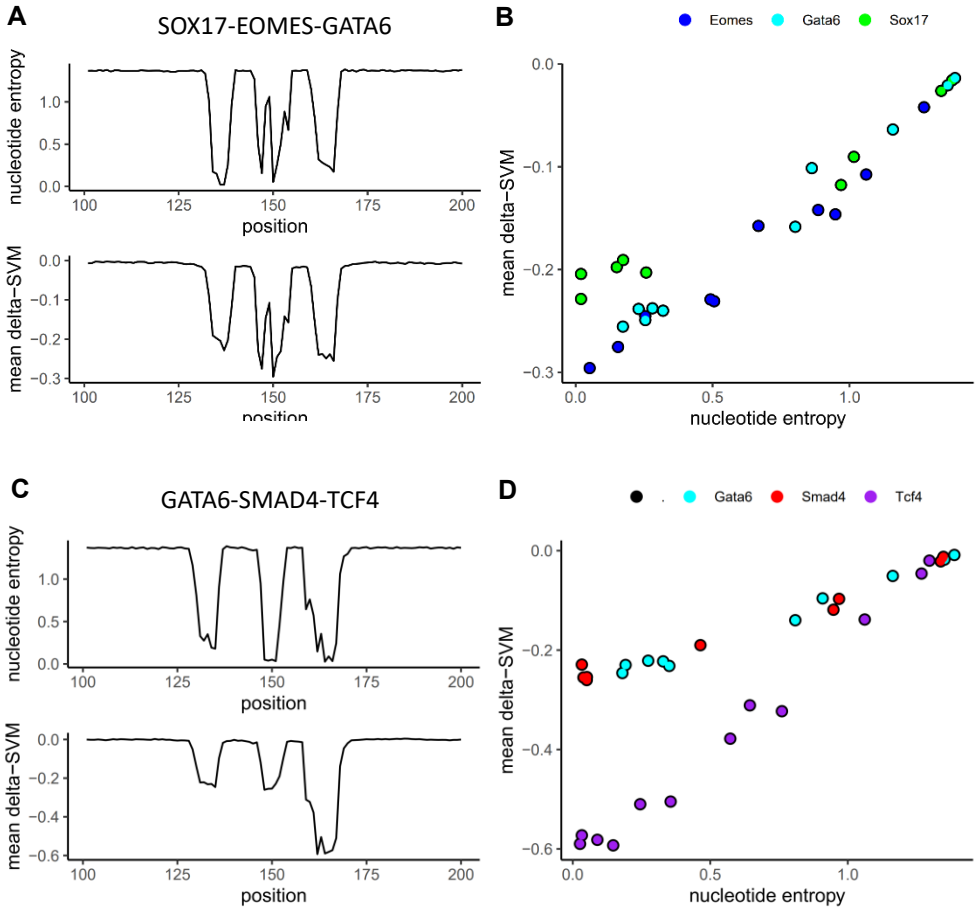

Figure 4.2. delta-SVM prediction score encodes sequence degeneracy within TFBS

Nucleotide entropy (computed across 1,000 simulated sequences) and average mean delta-SVM (N=1,000) for each base pair positions for **AB)** type 1 enhancer (SOX17-EOMES-GATA6) and **CD)** type 10 enhancer (GATA6-SMAD4-TCF4). **BD)** plotting nucleotide entropy vs mean delta-SVM in **A** and **C** for motif-seeded regions.

and computed mean delta-SVM scores for each enhancer (**Methods 4.2.1, 4.2.2**). **Figure 4.1B** shows that the mean delta-SVM scores, averaged within types, recover the base pairs where each TF-binding motifs were seeded, while delta-SVM scores at base pairs with no motif seeding (i.e. genomic background) are small and centered around zero. Hence, the simulation shows that gkm-SVM kmer-weights can encode sequence patterns specific to the seeded motifs.

Further, kmer-weights encode sequence degeneracy within TFBS (**Figure 4.2**). For each of the 10 enhancer types, I computed nucleotide entropy at each base pair position, and compared the nucleotide entropy with average delta-SVM predictions. Sequence degeneracy within a motif site was quantified with nucleotide entropy, and the nucleotide entropy showed high correlation with average delta-SVM scores. **Figure 4.2AB** and **Figure 4.2CD** each shows the result for the type 1 enhancer (SOX17-EOMES-GATA6) and the type 10 enhancer (GATA6, SMAD4, TCF4).

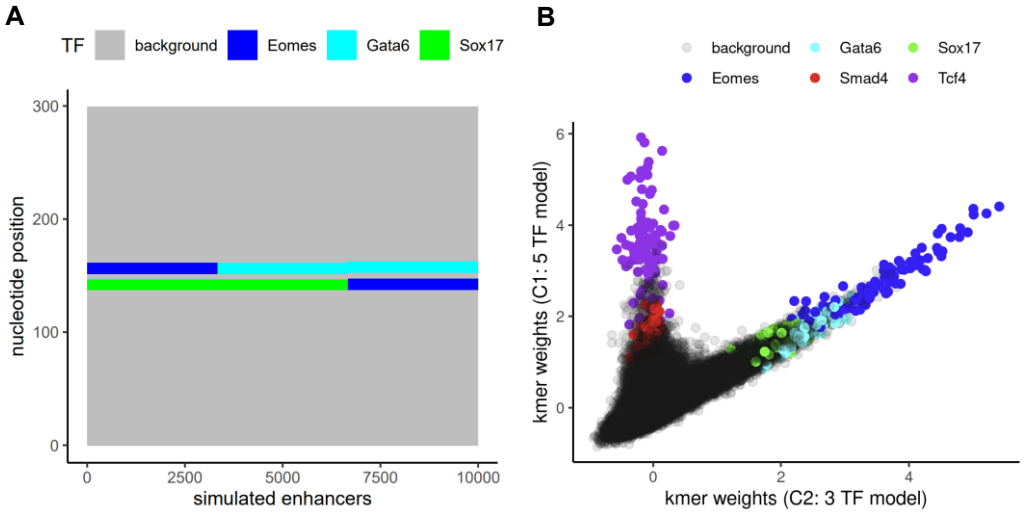

Figure 4.3. gkm-SVM kmer-weights capture differential motif enrichment in distinct enhancer sets.

A) Visualizing 3 classes of synthetic enhancers by seeded PWM TF motifs. Total 10,000 enhancers (x-axis) of width 300bp (y-axis) were simulated. Each base pair position is colored by seeded PWM TF motif. **B)** plotting kmer-weights of two gkm-SVM models trained on the enhancer set visualized in Figure 4.1A or 4.3A. Kmers with significant sequence association with a TF PWM are colored accordingly.

Lastly, kmer-weights encode sequence patterns enriched in positive training set relative to the background negative set, and the weights can be compared to identify cell-specific core TF regulators. Let C_1 and C_2 be two hypothetical cell types, where C_1 and C_2 respectively has the synthetic sequences described in **Figure 4.1A** and **Figure 4.3A** as their cis-regulatory modules. By design, the enhancer sequences of C_1have binding sites for SOX17, EOMES, GATA6, SMAD4 and TCF4, but C_2 lacks binding sites for SMAD4 and TCF4. This difference in seeded TFBS distributions in C_1 and $C_2$are reflected in the comparison of their kmer-weights (**Figure 4.3B**). Kmers associated with the three common TFs (SOX17, EOMES, GATA6) have high weights in both kmer-weights, while kmers associated with C_1-specific TFs (SMAD4, TCF4) have high weights only in C_1 (**Methods 4.2.3**).

4.4 Modeling human and mouse enhancers with gkm-SVM

We generated sequence models of enhancer sequences of diverse human and mouse cell/tissues to identify regulatory vocabularies that drive cell-specific enhancer activities. Majority of the models were generated using DNase-seq data generated by various ENCODE consortium labs[32,97,100,141], and these models are publicly accessible through the ENCODE portal. In this section, I explain how we generated the gkm-SVM models and used the models to quantify regulatory conservation between human and mouse. This section is adapted from the *gkm-align* manuscript[10] [5].

Enhancers are distal to transcription start sites (TSS) and harbor binding sites for transcription factors (TF). The cell-specific expression of these TFs leads to cell-specific

[5] All the supplementary tables and other relevant information are accessible through beerlab.org/gkmalign/.

enhancer chromatin accessibility and transcriptional regulation. Over the past decades, the ENCODE consortium has generated thousands of DNase-seq experiments[97,100,141,142], across diverse human and mouse cell/tissues, and this comprehensive library of experiments has allowed us to robustly and systematically identify enhancer elements. Chromatin accessible peaks broadly fall into enhancer and promoter classes, which have different sequence features and conservation properties. To robustly define these classes, we use two genomic features: distance to the nearest TSS, and the cell-specificity of DNase I hypersensitivity (**Figure 4.4A**). We quantified the cell-specificity of a DNase I hypersensitive site (DHS) as the fraction of all biosamples in

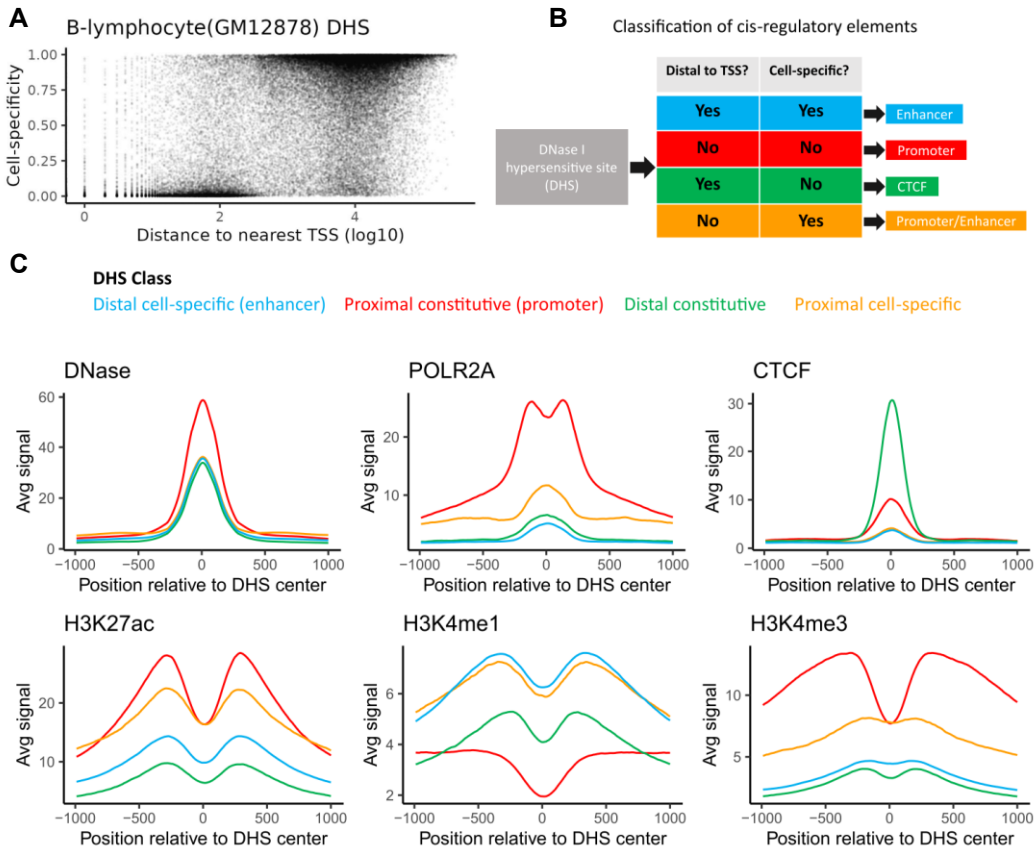

Figure 4.4. Defining enhancers as cell-specific distal DHS

A) Classifying DHSs by cell-specificity and TSS distance. **B)** B-lymphocyte DHSs (N = 55,715) distance to nearest TSS and cell-specificity across 1,270 biosamples. **C)** Classification of B-lymphocyte DHSs by their distances to nearest TSS (proximal: < 2kb, distal: > 2kb) and cell-specificity (cell-specific: DHS in less than 30% of all biosamples, constitutive: otherwise); average epigenetic signals around DHS peak centers by DHS class. Signals normalized as fold-change over genomic average.

the ENCODE database in which the DHS is *inaccessible* (N_{human}=1,270; N_{mouse}= 153). To discretize these classes, we classified all DHSs farther than 2 kilobases from the nearest TSS as *distal* (if not, *proximal*) and DHSs with cell-specificity higher than 0.7 as *cell-specific* (if not, *constitutive*) (**Figure 4.4B**). This partitions DHSs into four classes: distal cell-specific, proximal constitutive, distal constitutive, and proximal cell-specific DHSs (**Methods 4.2.4**).

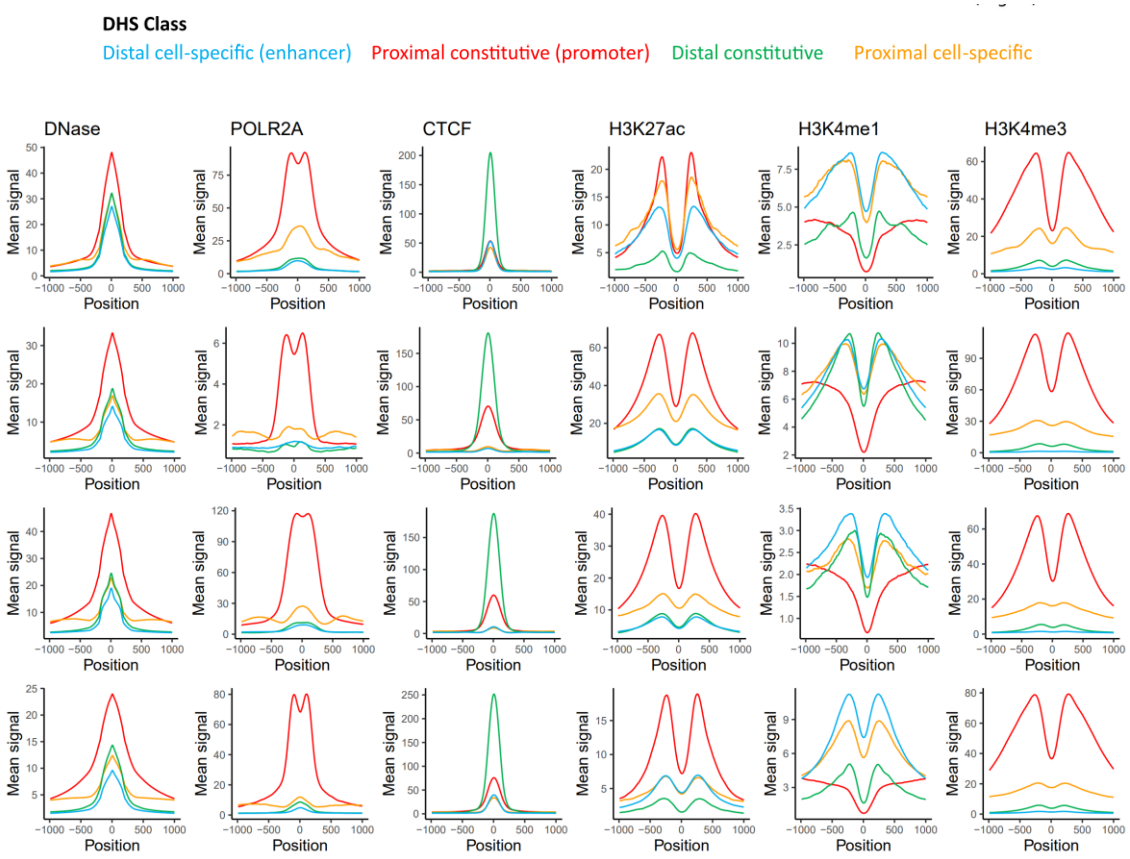

Figure 4.5. Epigenetic profiles of DHS subclasses across diverse human cell/tissues

Top to bottom row: ESC/adult heart/adult liver/ neural progenitor. Classification of DHSs from diverse cell/tissues by their distances to nearest TSS (proximal: < 2kb; distal: > 2kb) and cell-specificities of chromatin accessibilities (cell-specific: DHS in less than 30% of all biosamples; constitutive: otherwise) and visualization of average epigenetic signals around DHS peak centers by DHS class. Signals normalized as fold-change over genomic average.

We observed distinct biochemical signatures in these four element classes. Distal cell-specific DHSs show strong markers of enhancer activity such as ChIP-seq signal for H3K4me1 and lack markers of promoter activity (POLR2A, H3K4me3) (**Figure 4.4C, Figure 4.5**) (**Methods 4.2.5**; ENCODE experiment file IDs listed in **Supplementary Table 1**). By contrast, proximal constitutive DHSs have the highest level of chromatin accessibility among the four classes and display clear signatures of promoter activity (POLR2A, H3K4me3) and depletion of enhancer marks (H3K4me1). These classification criteria allowed us to robustly define enhancer elements without the need for diverse histone ChIP-seq experiments, which are currently unavailable for many biosamples assayed with the DNase-seq experiments. Further, many distal constitutive DHSs appear to be CTCF binding sites (a known regulator of chromatin topological organization), and proximal cell-specific DHSs show mixed signatures of enhancers and promoters, which emphasizes the utility of the two criteria (TSS distance; cell-specificity of chromatin accessibility) for precisely defining enhancer elements. For the rest of the dissertation, I will refer to distal cell-specific DHSs as enhancers and proximal constitutive DHSs as promoters.

Enhancer regulatory vocabularies, obtained through gkm-SVM training on enhancers, are cell-specific. Gkm-SVM is a sequence-based machine learning method that learns to effectively distinguish enhancers from inactive genomic elements by learning the weighted combination of gapped-kmers that predicts enhancers[23]. Gkm-SVM training assigns higher weights to gapped-kmers enriched in enhancers, and this information can be mapped to kmer weights, where kmers comprised of predictive gapped-kmers – or enhancer vocabularies – are assigned higher weights. The biological relevance of enhancer regulatory vocabularies has been demonstrated by their utility in predicting functional impacts of enhancer sequence variants[3,4,28,30,47]. All human and

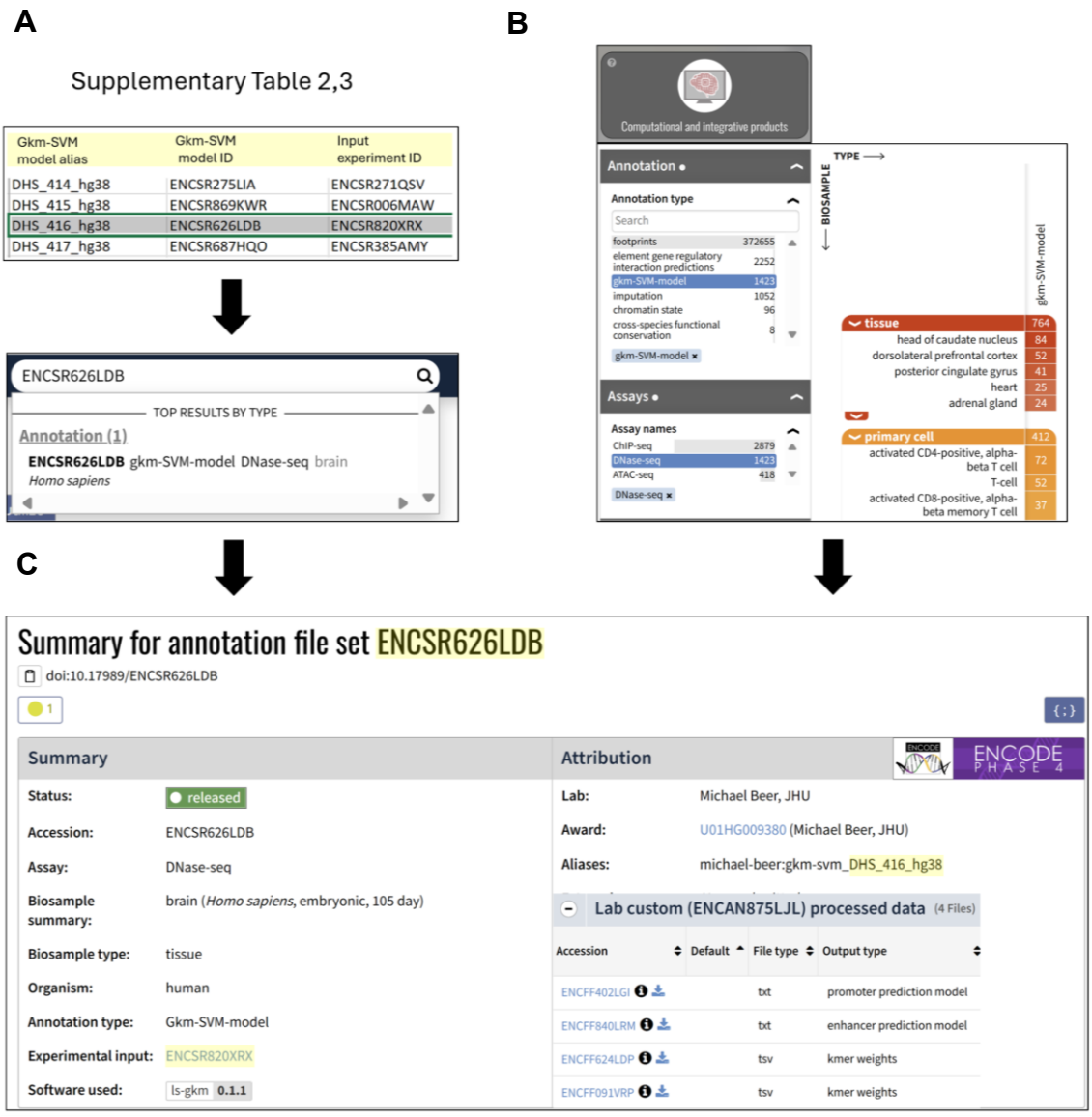

Figure 4.6. Accessing gkm-SVM models through the ENCODE portal

All the gkm-SVM models used in this study are publicly accessible through the ENCODE portal (encodeproject.org). **A)** These models can be individually searched using gkm-SVM model aliases, model IDs, or experiment IDs listed in Supplementary Table 2 (human) and Supplementary Table 3 (mouse). **B)** The full catalogue of gkm-SVM models that we made public (including those not used in this study such as ATAC-seq or ChIP-seq models) can be accessed through the “Computational and integrative products” icon in the main ENCODE webpage. **C)** gkm-SVM “Annotation file” for each DNase-seq experiment contains four files: 1. enhancer prediction model (trained using distal cell-specific DHS) 2. enhancer kmer weights 3. promoter prediction model (trained using proximal DHS) and 4. promoter kmer weights.

mouse gkm-SVM models used in this study are publicly available in the ENCODE portal (encodeproject.org), and their aliases and experimental inputs are listed in **Supplementary Table 2-3**. **Figure 4.6** describes how to access the gkm-SVM models through the portal.

Enhancer kmer-weight vectors encode DNA binding motifs of core TF regulators that determine cell/tissue identity, leading to similarities of enhancer regulatory vocabularies by cell/tissue types. For example, gkm-SVM models trained on human and mouse adult brain enhancers detect the same TFBS, encoded by highly similar enhancer regulatory vocabularies (**Figure 4.7;** R = 0.74). For example, one of the top predictive kmers (CACCAGATGGC) shared between human and mouse brains is a TFBS for neurogenic TFs NEUROG1/2 and ATOH1[140,143,144]. Another predictive kmer for both human and mouse brains (CCATGGCAACC) is bound by RFX family TFs, of which RFX3/ RFX5/ RFX7 are highly expressed in the human brain[145], and their mutations are linked to intellectual and behavioral abnormalities[146]. On the other hand, kmers associated with TFs highly expressed in non-neural cell/tissues, such as AAAGAGGAAGT (SPI family; blood cell development[147]) and GGACTTTGACC (HNF4A; liver/pancreas[148]), have low weights in both human and mouse brain enhancer models. Cell/tissues of distinct identityhave low similarity in enhancer kmer weight vector (**Figure 4.8**; $R_{\text{brain vs monocyte}}$, $R_{\text{spinal cord vs macrophage}}$, $R_{\text{brain vs liver}}$ = 0.085, 0.11, 0.19). Across a wider range of cell/tissues, biosamples of the same cell/tissue identity consistently show high kmer-weight correlation (mean R = 0.72) while pairs of kmer-weights from distinct cell/tissues show lower correlation (mean R = 0.31) (**Figure 4.9A**). By contrast, kmer-weights obtained from promoters show low cell-specificity, having high kmer-weight correlation for all pairs of cell/tissues (same tissue mean: 0.80; distinct tissue mean:

0.75). This is consistent with past observations that enhancers are bound by TFs with cell-specific expression while promoter binding TFs are relatively less cell-specific[24].

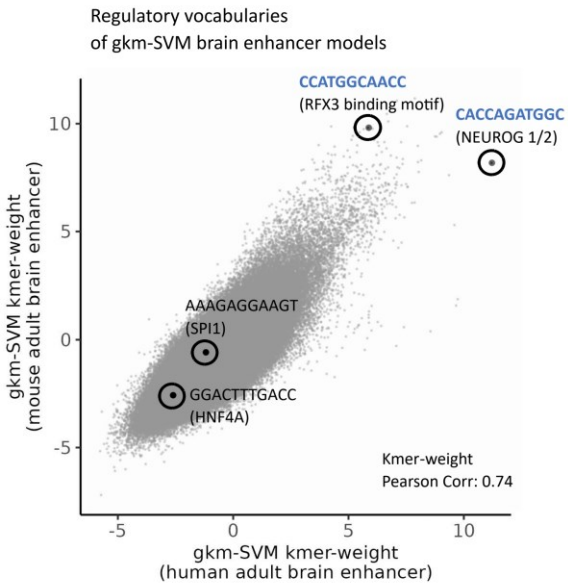

Figure 4.8. Enhancer regulatory vocabularies of human and mouse brains are conserved

Pairwise comparisons of kmer-weight vectors, derived from enhancers (cell-specific distal DHS), for a pair of similar samples: human and mouse adult brain (N_{kmer} = 2,097,152; Pearson Corr. = 0.74)

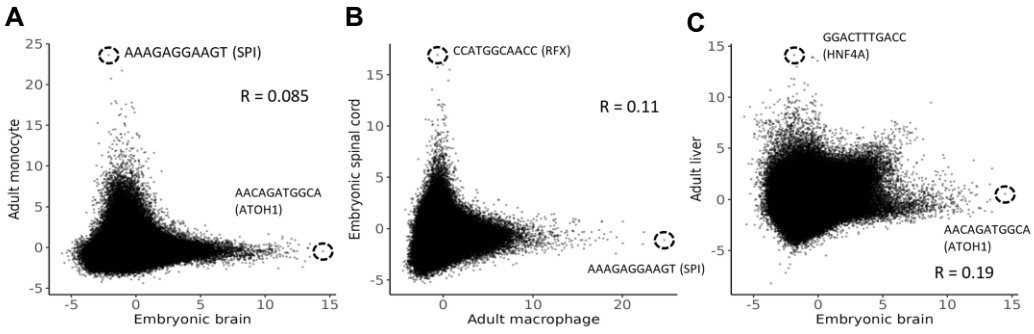

Figure 4.8. Enhancer vocabularies are distinct for distinct cell/tissue types.

Comparing kmer-weights between **A)** monocyte & brain **B)** spinal cord & macrophage **C)** brain & liver. Gkm-SVM models used for this figure can be identified by their model aliases and searched in the ENCODE portal (brain = DHS_399_hg38; spinal cord = DHS_421_hg38; monocyte = DHS_828_hg38; macrophage = DHS_482_hg38; liver = DHS_1229_hg38). More detailed information about these models is given in **Supplementary Table 2**.

Comparing all human and mouse tissue pairs, we find that enhancer regulatory vocabulary is conserved, as shown for a subset of cell/tissues in **Figure 4.9B**. We trained gkm-SVM enhancer prediction models for a set of human and mouse biosamples (brain, retina, B-cell, T-cell, intestine, stomach), and evaluated whether gkm-SVM regulatory vocabularies trained on one species are predictive of enhancers in the other species (i.e., train on human enhancers, predict on mouse enhancers & train on mouse, predict on human) (**Methods 4.2.6**). We evaluated the similarity of human and mouse models by calculating the average between reciprocal interspecies prediction accuracies (AUROC). For matched cell/tissues, the accuracy of interspecies enhancer prediction is high (N = 6; mean AUROC = 0.91), while the prediction accuracy between distinct cell/tissues is low (N = 30; mean AUROC = 0.65). On the other hand, interspecies prediction accuracies for promoters are high for both mismatched (e.g., human brain & mouse B-cell; N = 30; mean AUROC = 0.93) and matched cell/tissues (N = 6; mean AUROC= 0.94). These results are consistent with past observations that expression of core TFs[149,150] and their DNA binding affinities[151] are well-conserved between human and mouse.

Utilizing its cell-specific conservation, we used enhancer regulatory vocabulary to identify the most similar pairs of human and mouse cell/tissues for further quantitative assessment of conservation (**Chapter 5**). We generated a comprehensive list of 45 human/mouse cell/tissue pairs matched by gkm-SVM enhancer vocabularies (**Supplementary Table 4**; **Method 4.2.7**). We will also use this set of 45 pairs of cell/tissues to evaluate our novel genome-alignment method, *gkm-align*, against conventional methods (**Chapter 6**).

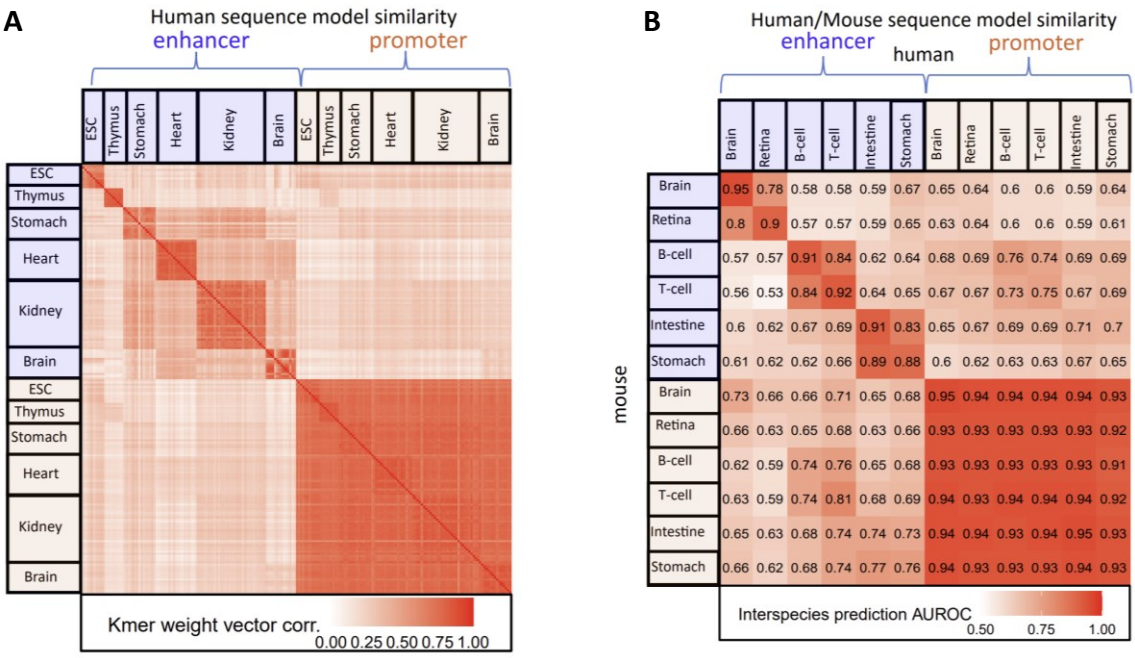

Figure 4.9. Similarity of enhancer vocabularies is cell-specific across diverse human and mouse tissues

A) Pairwise comparisons of gkm-SVM kmer-weight vectors, by Pearson correlation, across various human embryonic biosamples (N_{ESC} = 11 N_{Thymus} = 9, $N_{Stomach}$ = 15 N_{Heart} = 19, N_{Kideny} = 32, N_{Brain} = 14) trained on enhancers and promoters. **B)** Human-mouse interspecies DHS prediction accuracy (AUROC) across various cell-tissues (avg. between mouse element prediction using model trained on human elements and vice versa).

4.5 Discussion

Gene regulatory elements are composed of disordered clusters of degenerate TF binding sites, and deciphering their sequence structures is crucial for linking regulatory mutations to disruptions in gene expression and also to diseases. The sequence complexity of regulatory elements poses great challenges to modeling their sequence structures. In this chapter, I provided a detailed introduction to gkm-SVM, a SVM-based machine learning algorithm that effectively models regulatory sequences using gapped-kmer sequence features. Upon training on enhancer sequences, gkm-SVM learns to predict enhancers using weighted sum of gapped-kmer composition, and this information

can be mapped to a kmer-weight vector for more interpretable enhancer feature extraction, which I called the enhancer regulatory vocabulary. I demonstrated using synthetic enhancers that gkm-SVM regulatory vocabularies robustly extract the expected TF binding motifs (**Figure 4.3**). I then trained gkm-SVM on human and mouse enhancers across diverse cell/tissues, and showed that enhancer vocabularies are highly conserved between human and mouse in a cell-specific manner (**Figure 4.9**). On the other hand, promoter vocabularies are similar across distinct cell-tissues. This is consistent with previous observations that enhancer activities are cell-specific, while promoter activities are similar across distinct cell/tissues. Lastly, I utilized the cell-specific property of conservation in enhancer vocabularies to derive 45 pairs of human and mouse orthologous cell/tissues. This list will be used in the next two chapters to quantify enhancer evolution by cell-types (**chapter 5**) and evaluate the novel gkm-align algorithm for mapping conserved enhancers (**chapter 6**).

Chapter 5

Evolution and Conservation of Mammalian Enhancers

5.1 Introduction

Model organisms have been indispensable for understanding the functional roles of cis-regulatory elements (CRE) in complex biological phenomena such as fetal development and pathogenesis[6]. Some human CREs with putative functional roles have been validated by dissecting the mutational impact of their conserved orthologous counterparts in model animals[2]. However, identifying orthologous CREs, especially for distal enhancers, is computationally challenging both due to their rapid evolution and their degenerate sequence structures. Functional characterization of regulatory elements[152] and non-coding GWAS disease associated variants[13,18,153,154] typically begins with mapping human enhancers to mouse with sequence alignment, which suffers from low sensitivity, and thus the accurate identification of orthologous CREs has long been a bottleneck in efforts to improve our understanding of CRE function.

Enhancers are DNA sequences harboring multiple transcription factor binding sites (TFBS) and are important regulators of gene expression[155]. In spite of their importance, enhancers have evolved rapidly relative to protein-coding sequences[156,157] and promoters[20,141,158]. Since DNA motifs for TFBS are often degenerate and spacing between TFBS typically do not contribute to enhancer function, nucleotide substitutions can accumulate more readily without significant functional changes in enhancers[6], and this flexibility likely aids regulatory evolution. Further, duplication of redundant TFBS and

[6] Most of the contents in this chapter are adapted from the original gkm-align manuscript[10].

accumulation of enhancer mutations can lead to turnover of TFBS through stabilizing selection[6]. This flexibility also applies on a larger scale to combinations of enhancers within intergenic loci, as enhancer function is weakly constrained by position relative to the target promoter, and enhancers are often accompanied by redundant *shadow* enhancers that regulate the same target genes[8,15,16]. As a result, the functional and mechanistic properties of enhancers likely have facilitated their rapid turnover throughout evolutionary history while maintaining DNA binding specificities of TFs[141,151]. Our analysis below supports this picture of rapid enhancer evolution in the context of conserved TF binding specificity across a broad range of cell types.

Previously, several groups have observed that many putative enhancers, marked by chromatin accessibility[141], TF-binding[158], and/or histone modifications[20] characteristic of cis-regulatory elements (e.g., H3K27ac, H3K4me1/3), lack functional conservation at orthologous loci of distant mammals predicted by conventional genome-alignment (e.g., LASTZ[21,22]) and mapping algorithms (e.g., LiftOver[159]). This apparent lack of enhancer conservation is largely due to rapid evolution of distal enhancers, but limitations in conventional computational genome alignment algorithms to detect conservation can also contribute significantly. Most conventional genome alignment algorithms utilize a *seed-and-extend* strategy, where short sequence matches are first obtained as *seeds* and then *extended* from both ends for further base pair alignment[21,22]. However, such nucleotide-level modeling of enhancer evolution may not be optimal for resolving sequence structures of enhancers since enhancers are characterized by collections of multiple degenerate TFBS.

In this chapter, I describe our results from systematic quantification of cell/tissue-specific enhancer conservation, using the 45 orthologous human and mouse cell/tissues derived in **chapter 4**. Through multiple orthogonal analyses, we show that conservation

levels of enhancers depend strongly on the cell/tissue type, which is partly explainable by association with transposable elements (TE).

5.2 Methods

5.2.1 Defining orthologous syntenic intergenic loci in human and mouse

The list of all human-mouse orthologous protein-coding gene pairs (N = 15,712) was obtained from the mouse ENCODE consortium publication[100] (**Supplementary Table 5**). All the supplementary tables are accessible through beerlab.org/gkmalign/. 15,500 out of 15,712 gene pairs with their human and mouse gene ID's also present in the Ensembl database were used , and the remaining 212 genes were filtered out (Homo_sapiens.GRCh38.96.chr.gtf, Mus_musculus.GRCm38.96.chr.gtf). Coordinates of these conserved human and mouse 15,500 genes were extracted from the Ensembl gtf files.

To identify all human-mouse syntenic intergenic loci, we first identified all neighboring pairs of human protein-coding genes conserved in mouse. Denote such human gene pair as HG_1 and HG_2. If their mouse gene orthologs, MG_1 and MG_2, are also neighbors in the mouse genome and if the relative transcriptional directions of [HG_1 and HG_2] and [MG_1 and MG_2] are also preserved (i.e., HG_1/HG_2 and MG_1/MG_2 both have tandem, convergent, or divergent transcriptional directions), we label [HG_1/HG_2 , MG_1/MG_2] as human-mouse syntenic neighboring gene pairs (N = 12,455). Human *syntenic intergenic locus* was defined as the union of genomic space between a pair of human syntenic neighboring genes and their gene bodies, and mouse *syntenic intergenic locus* was defined similarly. This led to 12,455 pairs of human and mouse

genomic regions that we call “human-mouse syntenic intergenic loci” (**Supplementary Table 6**).

5.2.2 Using LASTZ/LiftOver for estimating conservation rate of cis-regulatory DNA elements

LASTZ (v1.03.66) chain files for human-mouse genome alignment were downloaded from these links (June 2020):

https://hgdownload.soe.ucsc.edu/goldenPath/hg38/vsMm10/hg38.mm10.all.chain.gz

https://hgdownload.soe.ucsc.edu/goldenPath/mm10/vsHg38/mm10.hg38.all.chain.gz

LiftOver software was downloaded from this link (June 2020):

https://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver

Minmatch=0.01 and -multiple options were used for high sensitivity and to allow duplicate mappings.

Command to map human elements to mouse:

liftOver h_elems.bed hg38.mm10.all.chain.gz h_elems_mapped.bed h_elems_not_mapped.bed -minMatch=0.01 -multiple

Command to map mouse elements to human:

liftOver m_elems.bed mm10.hg38.all.chain.gz m_elems_mapped.bed m_elems_not_mapped.bed -minMatch=0.01 -multiple

With these settings, LASTZ/LiftOver maps a human DHS (e.g., human brain enhancer) to 1≥ mouse locus, and we say it is *conserved* if at least one of the mouse mapped loci overlaps (≥1 bp) with a mouse DHS in the matched cell/tissue in the matched syntenic intergenic loci. Human DHS conservation rate is defined as the

fraction of conserved human elements, and mouse DHS conservation rate is computed similarly. Human-mouse conservation rate is defined as the average of the two rates.

5.2.3 Sequence homology analysis for quantifying the proportions of orthologous and paralogous enhancers

Our usage of the terms orthologous and paralogous enhancers is an estimate used to understand these observations, and is not meant to reflect an evolutionary process. Sequence similarity between a pair of enhancers was quantified as the similarity in gapped-kmer composition (*gkm-similarity*; Eq.4.2 in **chapter 4**) using gkmSVM[23] R software package (v0.81.0). Prior to computing gkm-similarities, we masked portions of enhancer sequences that are predicted to be highly prevalent genome-wide to prevent trivial sequence matches by ubiquitous sequence patterns such as the low-complexity repeats (**Figure 6.2**). About 10% of enhancer sequence base pairs were masked and replaced with random base pairs

To compute proportions of human orthologous and paralogous enhancers in each human/mouse cell/tissue pairs (e.g., human brain and mouse brain), we obtained top 5,000 enhancers with highest DNase I signals (MACS2 peak score) and computed their pairwise sequence similarity with all other 5,000 human enhancers and with every mouse enhancer of the matched cell/tissue type. If the number of mouse enhancers exceed 50,000, we used the top 50,000. Using these values, we identified top matched human and mouse enhancers that are most similar to each of the top 5,000 human enhancers. Sequence similarities with top matched human and mouse enhancers were then used to classify a human enhancer as orthologous, paralogous or neither. Denoting top sequence similarity with human enhancers and with mouse enhancers as y and x respectively, we classified human enhancers according to the following classification rules:

$$\begin{cases} \text{if } x, y < b & \text{no sequence homolog} \\ \text{else} \begin{cases} \text{if } y > ax + b(1-a) & \textbf{paralogous} \\ \text{if } y < \frac{1}{a}x + b\left(1 - \frac{1}{a}\right) & \textbf{orthologous} \\ \text{else} & \text{ambiguous} \end{cases} \end{cases}$$

, where (a, b) = (4, 0.1). The decision regions for orthologous and paralogous enhancers are each shaded blue and red in **Figure 5.6A**. Mouse enhancers were classified similarly.

5.2.4 Annotations for repetitive DNA elements

Annotations for repetitive elements were downloaded from the following links[160]:

http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/rmsk.txt.gz

http://hgdownload.cse.ucsc.edu/goldenpath/mm10/database/rmsk.txt.gz

5.3 Results

Although enhancer vocabularies are highly conserved, mapping orthologous enhancers conserved between distant mammals remains computationally challenging. To assess enhancer conservation rate, we will use as a metric the fraction of predicted enhancers in one species that is also chromatin accessible in the other species, quantifying enhancer conservation rate between two species as the average fraction of the two reciprocal directions (**Figure 5.1A**). To demonstrate the challenge of mapping enhancers, we will first use a conventional mapping method (LASTZ/LiftOver) and compute the conservation rate of human fibroblast enhancers and promoters using a set of DHS data in fibroblasts from diverse mammals[138] (chimpanzee, gorilla, orangutan, rhesus, and mouse[141], in increasing divergence from human) (**Method 4.2.4**; **Figure 5.1A**). About 40% of enhancers are mappable between human and chimpanzee, and

this value rapidly decreases to 11% in human and mouse as evolutionary distance to human increases (72% decrease in conservation rate; **Figure 5.1A**). Promoter conservation rate also decreases from 89% between human and chimpanzee to 65% between human and mouse but at slower rate of 27%. In contrast, regulatory vocabularies of both enhancers and promoters as quantified by gkm-SVM are constant across all species (**Figure 5.1A**); interspecies prediction accuracy (reciprocal average AUROC) of enhancers and promoters between human and mouse (AUROC enhancer:

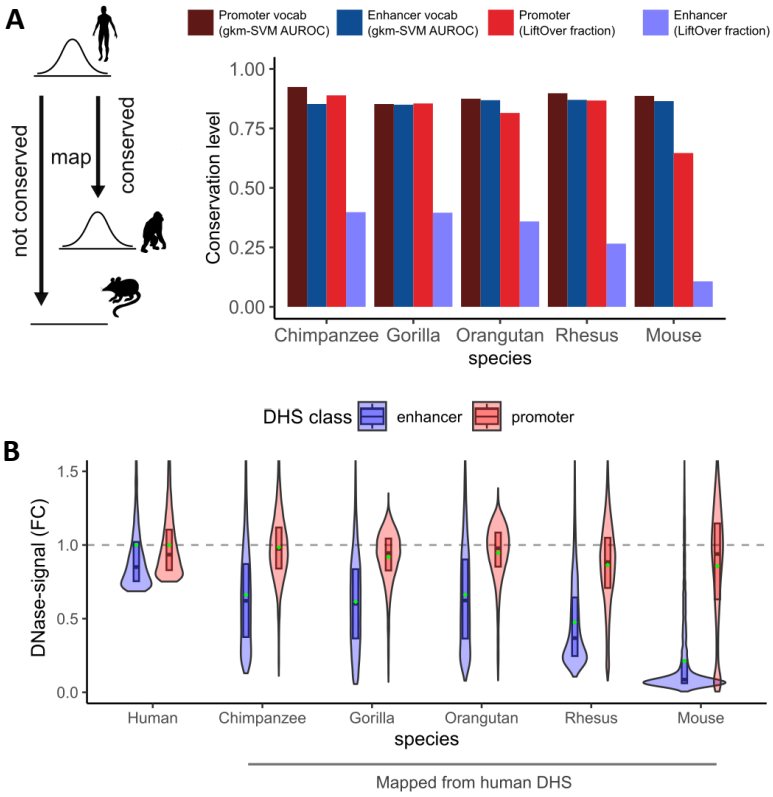

Figure 5.1. Enhancer and promoter regulatory vocabularies are conserved, yet enhancers rapidly evolve.

A) Schematic of enhancer mapping and conservation level of human fibroblast enhancer/promoter regulatory vocabularies (gkm-SVM interspecies enhancer prediction AUROC) and enhancer/promoter CREs (fraction of conserved enhancers mappable by LiftOver) in chimpanzee, gorilla, orangutan, rhesus, and mouse. **B)** Distribution of DNase accessibility in human fibroblast enhancers and promoters (top 10,000 in each DHS class by total DNase-seq reads mapped) and distribution of DNase-signal at primate and mouse loci mapped from human enhancers and promoters. Signals are normalized as fold changes from average fibroblast enhancers and promoter accessibilities in respective species (top 10,000 in each DHS class by DNase-seq read mapped).

0.86; promoter: 0.89) are as high as the prediction accuracy between human and chimpanzee (AUROC enhancer: 0.85; promoter: 0.92). As an alternative metric, we can count the read signal in the orthologous loci relative to average elements in that class. This confirms the rapid reduction in enhancer conservation rate (**Figure 5.1B**), where we show, for each species, distributions of DNase-signal at orthologous loci mapped from human fibroblast enhancers and promoters. Orthologous chimpanzee loci mapped from human enhancers and promoters respectively are on average 66% and 98% as accessible as average chimpanzee enhancers and promoters. These signals decrease dramatically to 21% (enhancer) and 86% (promoter) when we map human enhancers and promoters to mouse, further underscoring the rapid evolution of enhancers.

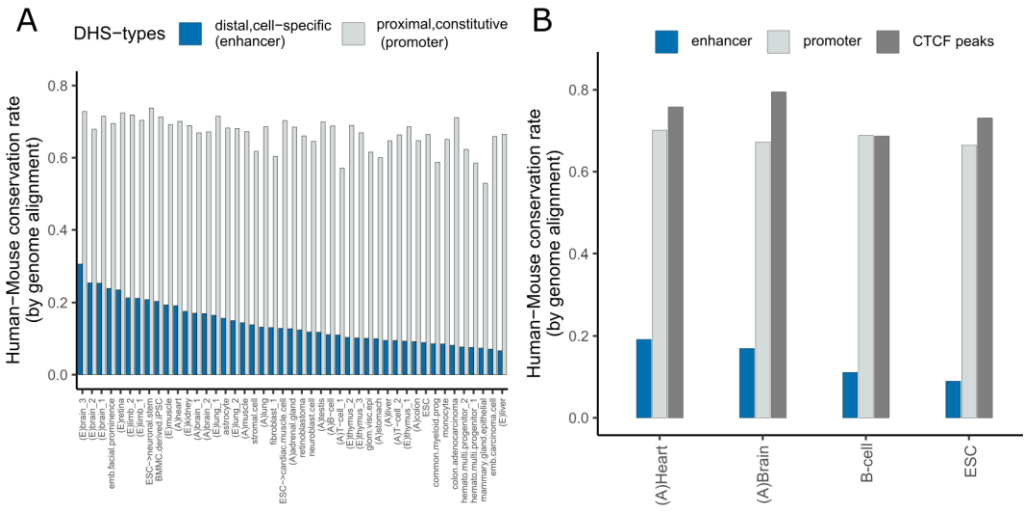

Figure 5.2. Enhancers are weakly conserved, and have cell-specific variation in enhancer conservation.

A) Human-mouse enhancer (distal cell-specific DHS) and promoter (proximal constitutive DHS) conservation rates (mappability to DHS by LASTZ/LiftOver genome alignment/mapping algorithms in orthologous syntenic intergenic loci. Mean of human to mouse and mouse to human mappings) across the 45 human-mouse cell-tissue pairs (A: adult, E: embryonic); sorted by enhancer mappability **B)** Human-mouse conservation rate for enhancers, promoters, and CTCF ChIP-seq peaks for B-cell, brain, ESC, and heart.

In the 45 pairs of human and mouse cell/tissues, we observed intriguingly high cell/tissue- specific variability in enhancer conservation (**Figure 5.2A**). As in **Figure 5.1A**, we defined the conservation rate of human enhancers as the fraction of human enhancers that map to mouse DHS of the matched cell/tissue (e.g., human brain/mouse brain), constraining the mapping by LASTZ/LiftOver to syntenic intergenic loci (**Method 5.2.1**). Conservation rates of mouse enhancers were defined similarly, and human-mouse conservation levels were defined as the average of the two reciprocal directions.

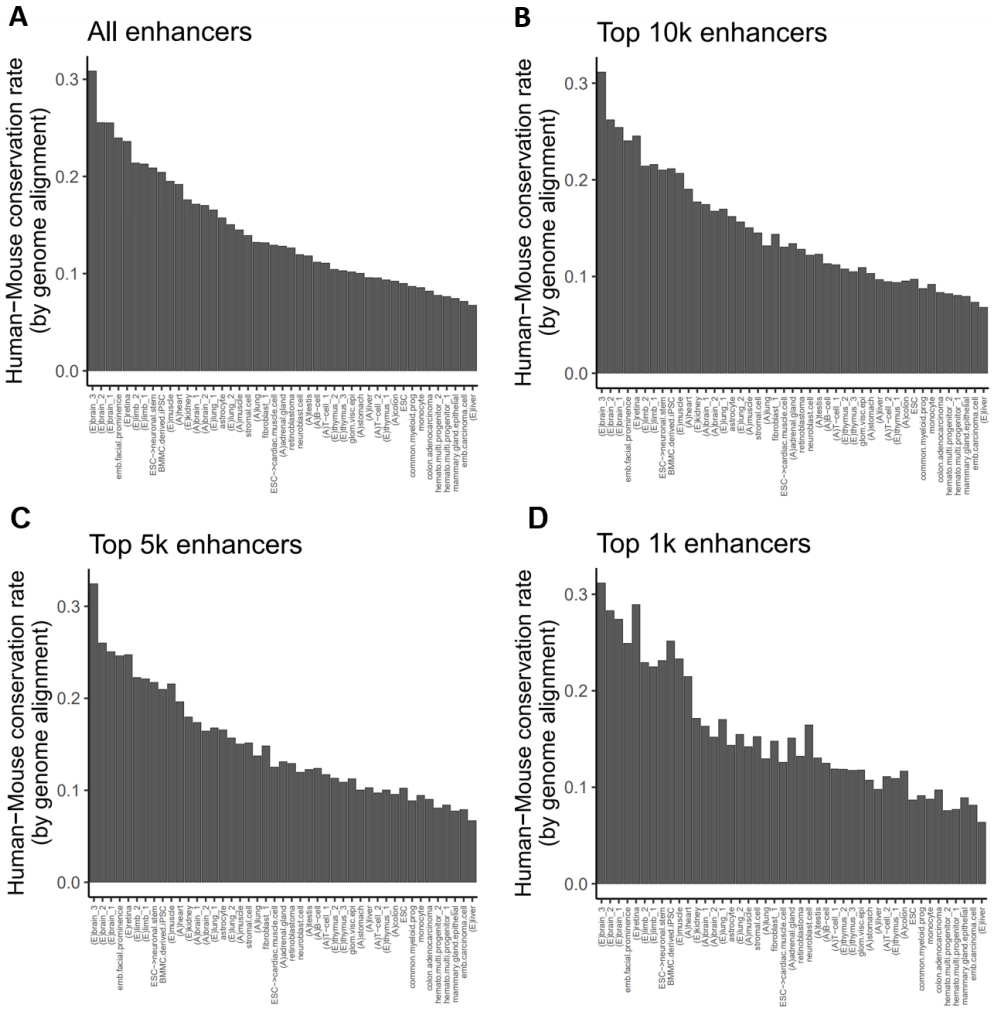

Figure 5.3. The cell-specific pattern of enhancer conservation is consistent across a wide range of enhancer depths.

A) same as the enhancer conservation rates shown in main figure 2A. **(B,C,D)** Similar to main Figure 2A but quantifying enhancer conservation rate as the fraction of top **B)** 10,000, **C)** 5,000, and **D)** 1,000 enhancers mappable to a DHS in the other species.

Promoters showed consistently high conservation rate across the 45 tissues (mean 67%) while enhancers showed highly variable conservation rate, ranging from 6.7% (embryonic liver) to 31% (embryonic brain) (**Figure 5.2A**). This cell-specific pattern of conservation persists even when we limit our quantification to enhancers with the highest DNase signals (**Figure 5.3**). Such lack of conservation is not observed in CTCF ChIP-seq peaks, which are also often distal to TSSs (conservation rate for brain, heart, B-cell, ESC = 82%, 79%, 72%, 76%; **Figure 5.2B**), suggesting that many CTCF loops and topologically associated domains are conserved[8,63,64,161]. The strong and somewhat counter-intuitive tissue specificity of enhancer conservation will be explored extensively below.

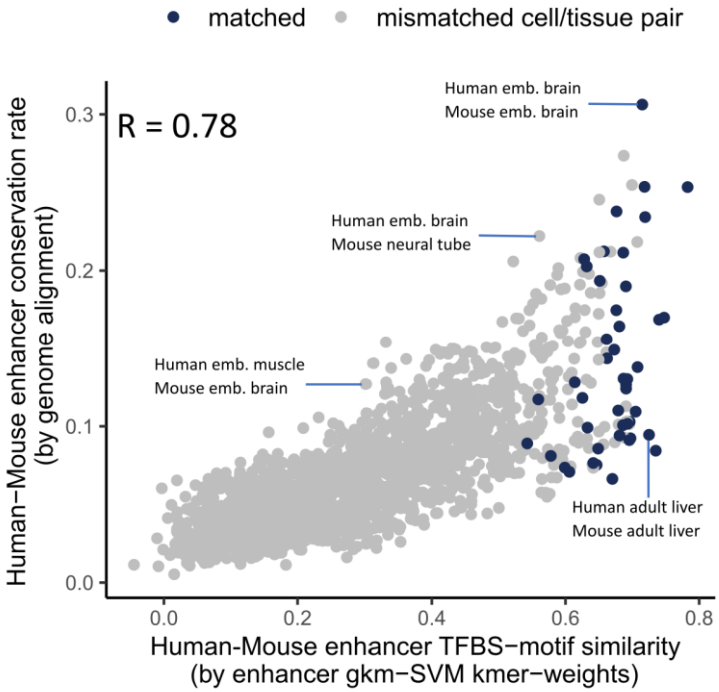

Figure 5.4. Comparing enhancer conservation and enhancer TFBS motif conservation

Human-mouse enhancer conservation rate by alignment (mappability by LASTZ/LiftOver) vs. human-mouse enhancer regulatory vocabulary conservation rate (correlation of gkm-SVM kmer-weights) for pairs of orthologous gkm-SVM matched tissues (e.g., human and mouse brain; N = 45) and mismatched cell/tissues (e.g. human T-cell and mouse brain; N = 1,980) (R =.78)

We observed low enhancer conservation rates in some tissues in spite of the fact that their regulatory vocabulary is highly similar between human and mouse. To explore this systematically, we computed the similarity of enhancer regulatory vocabularies (measured as Pearson Corr. of enhancer kmer-weight vectors), for both matched and mismatched pairs of human and mouse cell/tissue pairs (e.g., matched: human brain & mouse brain; mismatched: human brain & mouse muscle; $N_{matched}$ =45; $N_{mismatched}$ = 45^2-45=1,980; **Figure 5.4**). Overall, conservation rates of enhancers and similarity of enhancer regulatory vocabularies correlate highly (R = 0.78, **Figure 5.4**), indicating that human and mouse cell/tissues that share similar sets of core TF regulators also tend to share a higher number of orthologous enhancers. While enhancer regulatory vocabularies were overall highly similar for all matched human/mouse cell/tissue pairs (relative to the mismatched pairs), interestingly, their enhancer conservation rates varied widely (vertical spread of black points). For example, the human/mouse adult liver pair had almost identical similarity of enhancer regulatory vocabulary (R = 0.72) as human/mouse embryonic brains (R = 0.71), but the adult liver enhancer conservation rate was 9.4% (less than 1/3 of the embryonic brain), even lower than the enhancer conservation rate between a mismatched pair of human embryonic muscle and mouse embryonic brain (12.7%). This implies that some cell/tissues, while maintaining their core TF regulators and their DNA binding specificities, have experienced more incidence of enhancer turnover than other cell/tissues.

To eliminate the possibility that the highly cell/tissue specific rate of enhancer conservation is a bias of the LASTZ/LiftOver alignment/mapping algorithms, we performed an orthogonal analysis. Using 12,455 syntenic intergenic loci of human and mouse derived from 15,500 orthologous protein coding genes[100] (**Supplementary Table 5-6**), we simply counted the number of human and mouse enhancers located in each of

the matched syntenic intergenic loci (**Figure 5.5A**; **Method 5.2.1**). We compare the correlation between the number of human and mouse enhancers in respective syntenic intergenic loci, which we will refer to as "syntenic enhancer number constraint," and it imposes an upper limit for mappability of human/mouse enhancers in the matched syntenic loci. If the number of enhancers in syntenic intervals is not conserved, there is no way the enhancers can be conserved at the sequence level unless they arose through duplication. Embryonic human/mouse brain, which had the highest rate of enhancer conservation, also showed the highest level of syntenic enhancer number constraint (R=0.93; **Figure 5.5A**), while adult human/mouse liver had lower syntenic enhancer number constraint (R=0.67), with occasionally drastically different numbers of enhancers in matched syntenic intergenic loci. This is consistent with reports of species-

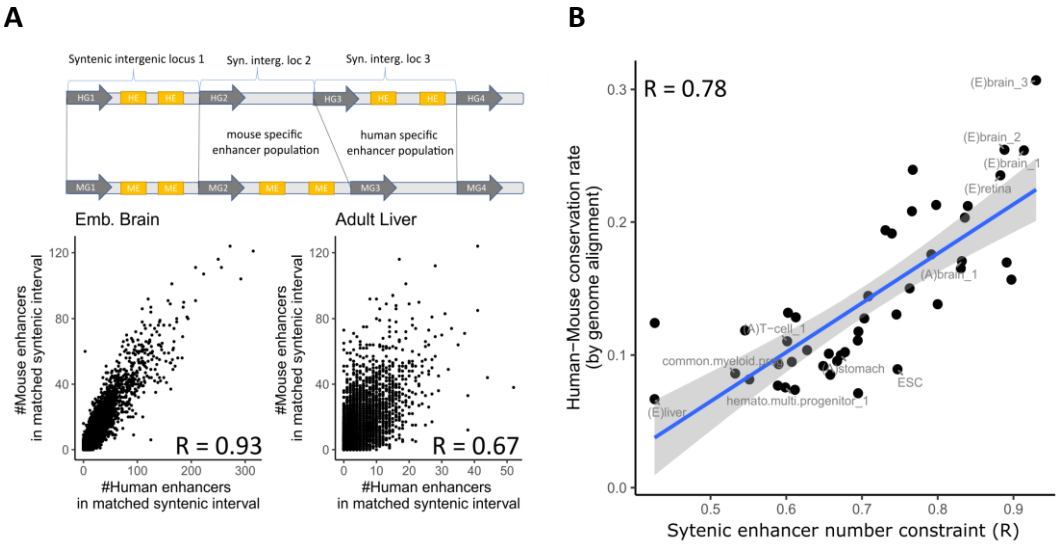

Figure 5.5. Enhancer conservation correlates with syntenic enhancer number constraint.

A) Schematics of human-mouse syntenic intergenic loci (HE: human enhancer; ME: mouse enhancer; HG: human gene; MG: mouse gene); Comparison of human and mouse enhancer numbers in matched syntenic intergenic loci (N = 12,704) for embryonic brain (syntenic enhancer number constraint R = .91) and adult liver (R = .67) **B)** Syntenic enhancer number constraint vs. human-mouse enhancer conservation rate (N = 45, R = .78; linear regression line and 95% confidence interval).

specific rewiring of transcription in the liver[162,163] and a relatively slower rate of transcriptomic divergence of the brain across mammals[25]. The lack of syntenic enhancer number constraint appears in cell/tissues with low enhancer conservation level (**Figure 5.5B**), and hints that the lower rate of conserved enhancers in some cell/tissues, as predicted by genome alignment, is an inherent property of the regulatory landscape.

This apparent lack of syntenic enhancer number constraint is largely driven by species-specific enhancer duplication, where cell/tissues with lower level of enhancer conservation tend to have a higher proportion of paralogous enhancers (see **Method 5.2.3**). To estimate the proportion of orthologous and paralogous enhancers in each human cell/tissue (in the context of human-mouse common ancestry), we labeled a human enhancer as paralogous if it has high sequence homology with another human

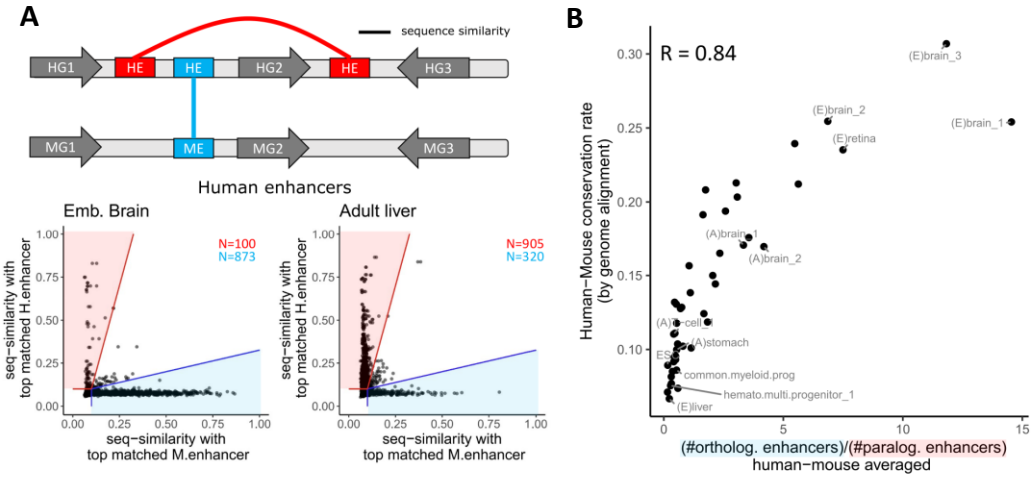

Figure 5.6. Cell/tissues with weak enhancer conservation are enriched with paralogous enhancers

A) Schematic describing how orthologous and paralogous enhancers are defined (HE: human enhancer, ME: mouse enhancer, G: gene; line indicates sequence homology); dots represent 5,000 human enhancers w/ highest DNase signal. Gapped-kmer sequence similarity with top matched mouse enhancers vs with top matched human enhancers. Enhancers in red shaded region are classified as paralogous enhancers; enhancers in blue shaded regions are classified as orthologous enhancers. **B)** Ratio of orthologous to paralogous enhancers across the 45 cell/tissue pairs (avg. of human and mouse) vs. human-mouse enhancer conservation rate (by genome alignment)

enhancer but lacks sequence homology with any mouse enhancer, and similarly labeled it as orthologous if it has high sequence homology with a mouse enhancer but lacks homology with any other human enhancers (**Figure 5.6A**). Based on this criterion, of the 5,000 human embryonic brain enhancers with highest DNase I accessibility, 873 and 100 were identified as orthologous and paralogous enhancers (ratio: 8.73). In contrast, adult liver had 320 and 905 orthologous and paralogous enhancers (ratio: 0.35). These estimated ratios of orthologous to paralogous enhancers, averaged between human and mouse, closely matched with the syntenic enhancer number constraint and with enhancer conservation rate for the 45 cell/tissue pairs (**Figure 5.6B**), indicating that enhancer duplication events have been a significant contributor to the divergence in enhancer landscapes.

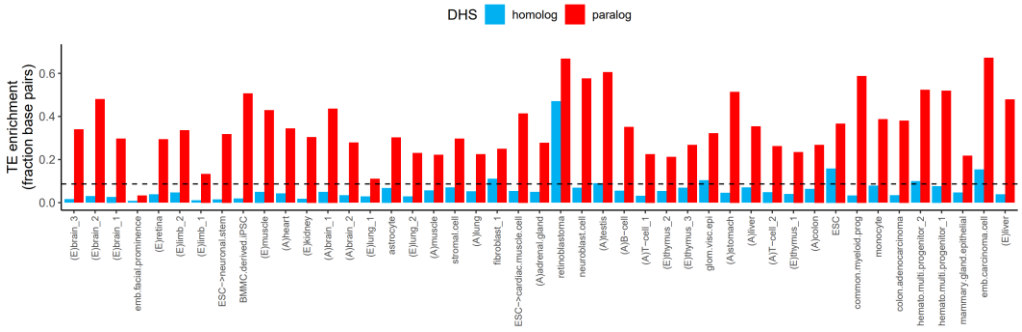

Figure 5.7. LTR transposable elements are enriched in paralogous enhancers.

Fraction of base pairs annotated as LTR transposable element for human enhancers classified as homologous or paralogous enhancers. The dotted line indicates the fraction of human genome annotated as LTR transposable element (0.087).

These duplication events are largely driven by transposable elements (TE). The paralogous enhancers show significant enrichment of LTR retrotransposons across diverse cell/tissues (**Figure 5.7**; **Method 5.2.4**), and we observed a general trend that the cell/tissue pairs with low enhancer conservation level tend to have high enrichment of transposable elements (**Figure 5.8A** – bars in the same order as **Figure 5.2A**; **Figure 5.11A**; R = -0.88). Quantifying TE enrichment as the fraction of total enhancer base pairs

that overlap with a TE, we observed that embryonic brain enhancers, averaged between human and mouse, had less than 10% enrichment of type I transposable elements (LTR/LINE/SINE) while some cell/tissues, such as ESC, had TE enrichment as high as 33%. Overall, TE enrichment in enhancers appears lower than its genome-wide coverage (>40%), with SINE elements especially depleted in enhancers of most cell/tissues (**Figure 5.8C**). However, interestingly LTR elements appear to be highly

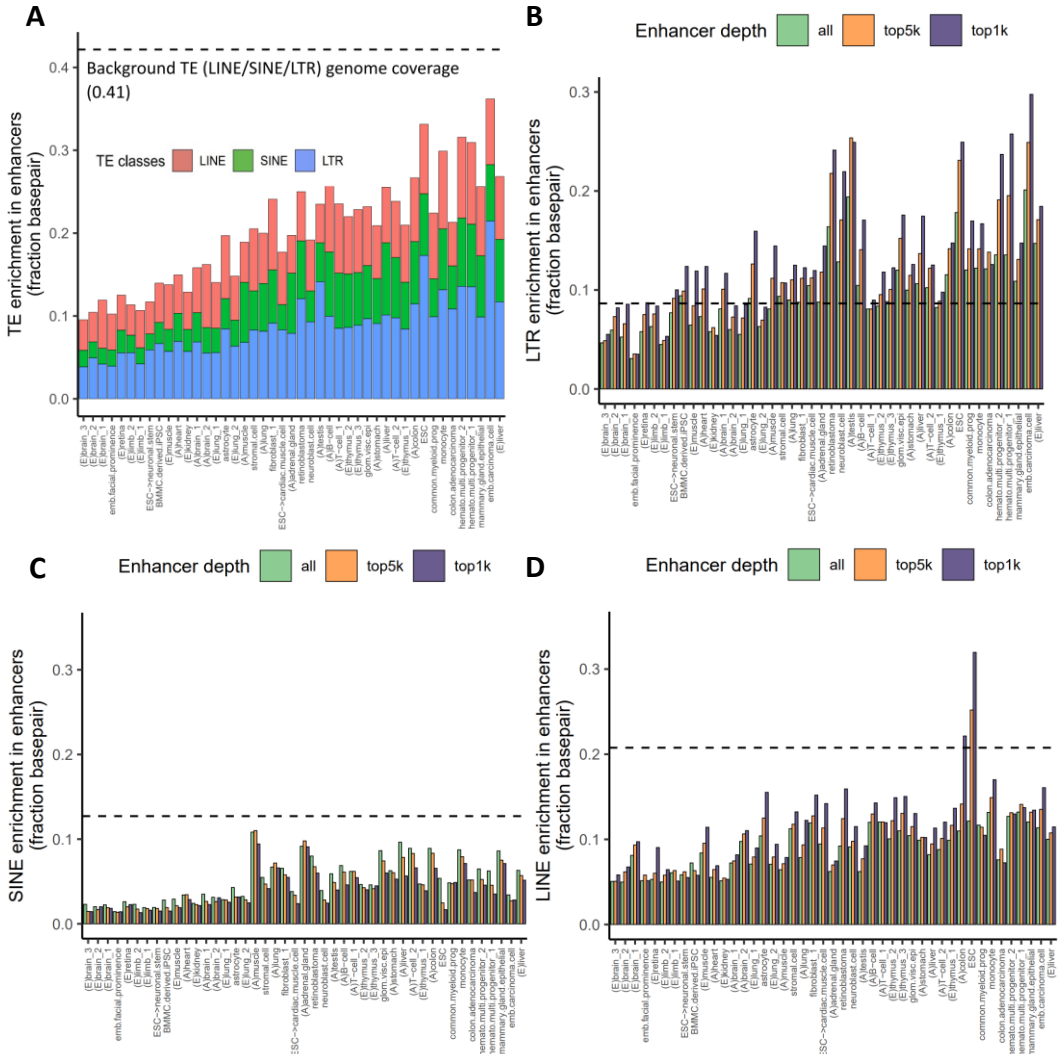

Figure 5.8. Weak enhancer conservation is associated with transposable elements.

A) Total fraction of enhancer base pairs annotated by each class of transposable elements (avg. of human and mouse). Fraction of human enhancer base pairs that overlap with **B)** LTR **C)** SINE **D)** LINE transposable elements across the 45 cell/tissues. The bar plots also show TE-enrichment for top 5,000 and 1,000 enhancers with highest DNase signal in each cell/tissue pairs. The dotted lines in **(A-D)** indicate genome-wide coverage (0.41, 0.087, 0.13,0.21). Identical tissue orders as Fig 5.2A.

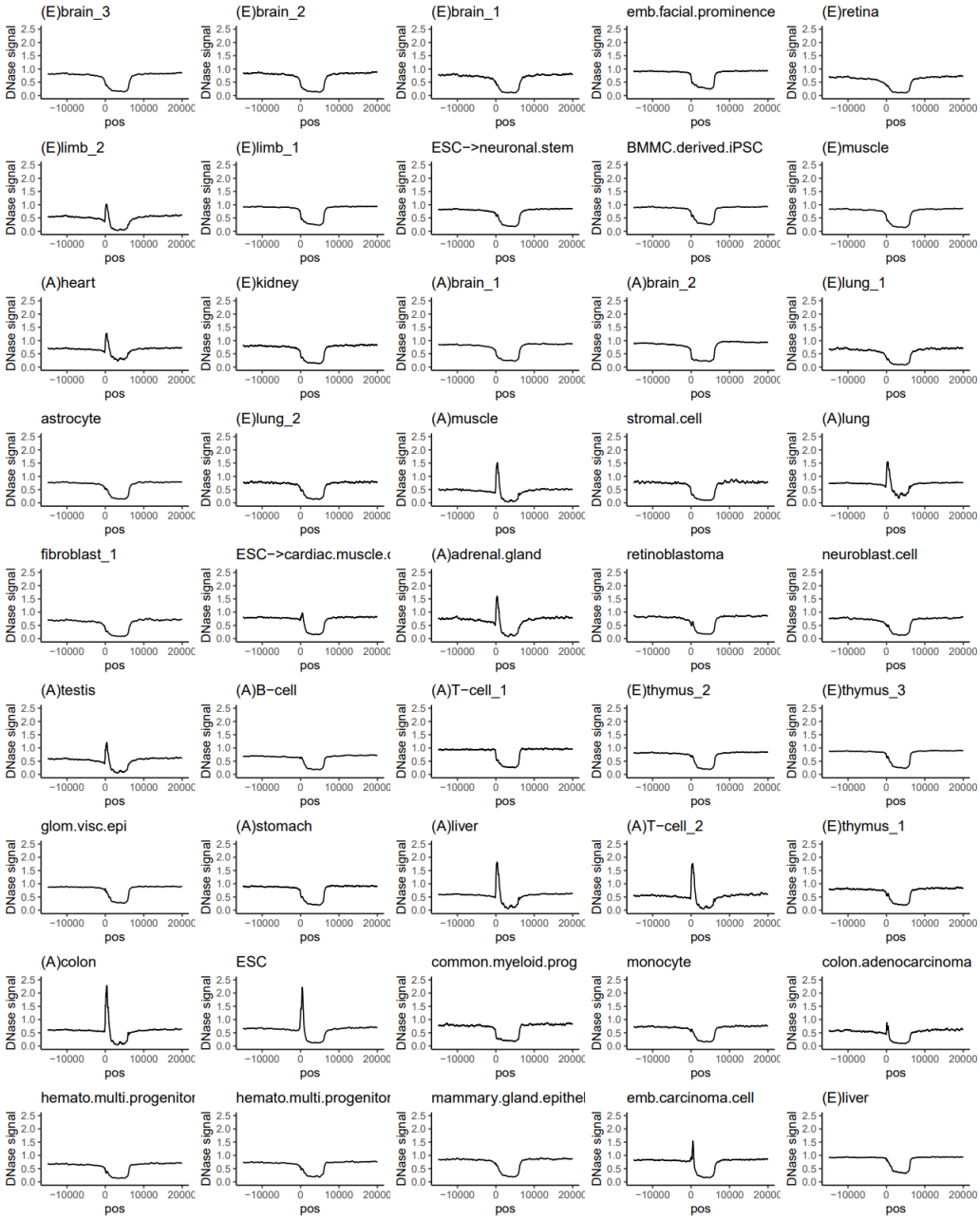

Figure 5.9. Average DNase accessibility profiles of full length human LINE1 elements

Average DNase accessibility profiles of full length (5,000 < size < 7,000) human LINE1 elements across the 45 cell/tissues. Position is relative to LINE1 5' end along their transcriptional directions. All cell/tissues show depletion of accessibility along LINE1 bodies while a subset of cell/tissues show high accessibility at 5' end (TSS). DNase signals computed as fold changes from genomic average.

enriched in enhancers across multiple cell/tissues, and their enrichment grows in enhancers with the strongest DNase I accessibility (**Figure 5.8B**). Although we do observe clear signals of DNase I accessibility in LINE elements (at its 5' end) for multiple cell/tissues (**Figure 5.9**), these generate weak DNase I peaks, and LINE elements are depleted in enhancers of most cell/tissues (**Figure 5.8D**). Like LTR, LINE enrichment increases with increasing DNase I accessibility, surpassing the genomic average for subsets of top 1,000 enhancers of human colon and ESC with the highest level of DNase I accessibility (**Figure 5.8D**). By contrast, SINE elements are more depleted in enhancers with higher DNase-I accessibility (**Figure 5.8C**). These observations are consistent with a recent report that also found significantly higher enrichment of LTR than SINE in distal enhancers[164]. This TE-specific and cell/tissue-specific variation in TE-enhancer association suggests that TEs may have a functional role in shaping theenhancer landscape[27,165–168], but it is difficult to separate function from their naturally increased tendency to transpose into accessible regions.

Most enhancers with overlapping TE annotations are species-specific (**Figure 5.10A**), and the increase in TE enrichment in enhancers explains much of the decrease in enhancer conservation across cell/tissues. This is not limited to weaker DHS, and in fact TE enrichment in enhancers grows with DNase-I accessibility. We show this by repeating the analysis of **Figure 5.10A** using only the top 1,000 enhancers with the highest DNase-signal (**Figure 5.10B**). In summary, enrichment of TE-associated enhancers contributes heavily to the observed cell/tissue dependent variability in enhancer conservation (orange bars in **Figure 5.10** grow with decreasing conservation). However, intriguingly, we also observe a decreasing trend of conservation in non-TE associated enhancers as TE-enrichment in cell/tissues increases (grey bars in **Figure 5.10** shrink with decreasing conservation; **Figure 5.11B;** R = -0.84). We find it curious

that the increased enhancer duplication driven by TEs is correlated with the reduction in the conservation level of non-TE associated enhancers. It is possible that TE driven functional redundancy allows more rapid evolutionary turnover in these tissues.

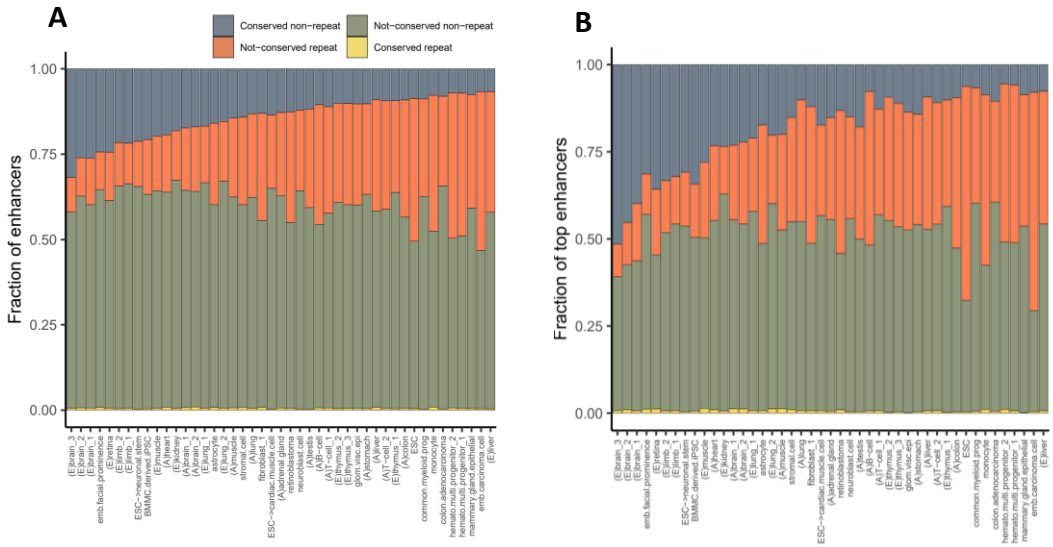

Figure 5.10. Enhancers overlapping transposable elements are species-specific.

A) Fraction of conserved enhancers and overlap with TEs (repeat = enhancers with more than 50% base pair overlap with type I TEs; non-repeat = enhancers with zero overlap with any repeat annotations). **B)** same but using top 1,000 enhance by DNase-signal.

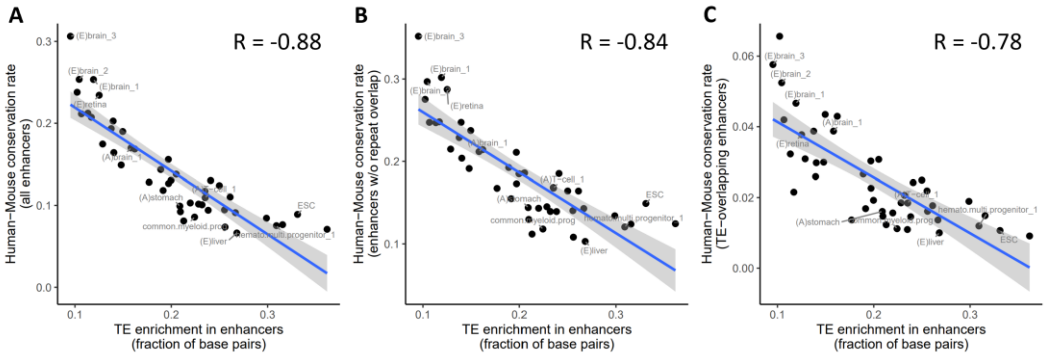

Figure 5.11. Human-mouse enhancer conservation rate decreases with TE-association.

A) Cell/tissues with high TE enrichment in enhancers have lower enhancer conservation rate (R = -0.88), which holds true for both subsets of enhancers **B)** with no repeat annotation (R = -0.84) **C)** with more than 50% overlap with LTR, LINE or SINE annotations (R = -0.78).

5.4 Discussion

Model organism studies have elucidated many of the mechanisms of transcriptional regulation and have functionally validated the roles of some enhancers associated with human diseases. However, characterization of enhancers through model animals is possible only for enhancers with identifiable orthologs in model animals. Mice are perhaps the most facile model system for human disease, but their evolutionary distance poses more of a challenge than detecting orthologous regulatory sequence in primates. Past studies using conventional genome alignment and mapping algorithms (LASTZ/LiftOver) have shown that mapping orthologous enhancers with conserved regulatory activities is much more difficult than mapping promoters. We addressed whether the difficulty of enhancer mapping is a result of rapid enhancer evolution or of limitations in conventional genome alignment algorithms. To comprehensively quantify enhancer conservation, we used gkm-SVM to generate unbiased sets of enhancers for each pair of 45 human and mouse cell/tissues matched by core TF regulators. Interestingly, enhancers appear to have highly variable levels of conservation across different cell/tissue types. We show that the conservation rate of embryonic brain enhancers are about three times higher than that of adult liver, although enhancer vocabularies are highly conserved for both tissues. The tissue variability in enhancer conservation rate is confirmed in alignment-free analyses. Part of the explanation of this apparent paradox is that tissues with low enhancer conservation also have more species-specific enhancer duplication. A significant proportion of enhancer duplications appear to have resulted from transposable elements (TE), especially LTRs, and most TE-associated enhancers are species-specific, partly explaining the relative lack of enhancer conservation in cell/tissues with high TE activity, such as the liver. The cell/tissue variability in enhancer conservation aligns with previously reported cell/tissue

variable conservation of gene expression, which showed rapid transcriptomic divergence of the liver relative to the brain[25]. Rapid liver enhancer evolution may explain the apparent transcriptomic differences between human and mouse livers[162,163], and may further explain the previously reported transcriptional divergence of many cell/tissue specific genes[169]. Intriguingly, in cell/tissues with low enhancer conservation, enhancers with no sequence overlap with repetitive elements also showed reduced conservation. Our observation is consistent with a model of evolution in which TEs provide influx of novel TF binding sites through transposition and further facilitate turnover of nearby enhancers by supplying functional redundancy[26]. TEs can contribute up to 30% of the strongest DHS peaks in some tissues, but we do not know what proportion of these TFBS-carrying TEs act as transcriptional enhancers to relevant genes, and which might be regulatory noise. Only a small proportion of TE-enhancers have so far been functionally tested[168], but we expect to see more functional validation of these elements in near future due to advances in noncoding CRISPR-based screening methodologies[9,170].

Chapter 6

Gkm-align: an algorithm to map conserved distal enhancers using gapped-kmer sequence features

In this chapter, I present the novel genome-alignment method that incorporates gapped-kmer features to model sequence degeneracy of enhancers: *gkm-align*[7]. This feature choice is motivated by sequence modeling using gapped-kmer composition which have been shown[171] to effectively represent biological sequences, accurately predict cell-specific enhancers, and discover regulatory vocabularies associated with TF binding (gkm-SVM)[3,4,23] and protein motifs[172]. The effectiveness of this modeling agrees with the prevailing model that enhancers are defined by clusters of degenerate TFBS[3,4,23,30]. *Gkm-align* incorporates this idea and aligns human and mouse sequences by their gapped-kmer composition. Using enhancers of the 45 human/mouse cell/tissue pairs, we systematically evaluated the *gkm-align* algorithm, and discover thousands of novel conserved enhancers. Further, we show that the discovery rate of conserved enhancers can further be increased by incorporation of gkm-SVM derived cell-specific regulatory vocabularies, which we show are conserved between human and mouse.

6.1 Methods

6.1.1 Whole genome-alignment and conserved enhancer mapping using gkm-align

Gkm-align software is accessible through its github page: https://github.com/oh-jinwoo94/gkm-align. Sample command lines with minimal examples and detailed

[7] Most of the contents in this chapter are adapted from the original gkm-align manuscript[10].

instruction of how to use gkm-align for genome-alignment and conserved enhancer mapping are also included in the github page.

6.1.2 Quantifying enhancer strength of human loci mapped from mouse HBB enhancers using CRISPRi perturbation data.

HCR-FlowFISH CRISPRi data (K562 cell line; HBE1 expression perturbation as readout) were downloaded from the ENCODE portal (accession ID provided in **Supplementary Table 1**; beerlab.org/gkmalign/)[9,103]. Two biological replicates were used, where each replicate generates sgRNA sequence read counts for low and high expression sort bins. The following equation was used to compute CRISPRi effect size of each sgRNA.

$$\log_2 FC_i = \log_2 \left(\frac{1 + \left(\frac{L_i}{\text{mean(L)}} \right)}{1 + \left(\frac{\text{H}_\text{i}}{\text{mean(H)}} \right)} \right)$$

where L and H are each a vector encoding the number of reads for each sgRNA in low and high sort bins respectively. Normalization with mean underweight sgRNAs with low read counts[9]. Enhancer strength of putative human enhancers, mapped from mouse HBB enhancers using gkm-align, were computed as average log2FC of sgRNA target that overlap with each enhancer. Mouse enhancer coordinates were defined using mouse embryonic liver DHS (ENCFF578VRG; 300 base pair wide extended from the summit) that also overlap with GATA1 ChIP-seq peaks in mouse erythroblast (ENCFF676TDJ) within mm10/chr7:103851395-103883181.

6.1.3 Regression model for predicting functional conservation.

For predicting functional conservation of human enhancers in mouse (measured as DNase-signal at mouse loci mapped from human enhancers), we denote a set of human enhancers as $\{HE_1, HE_2, \ldots\}$ and denote mouse elements mapped from query HE by gkm-align as ME_i. Denoting DNase-signal at query HE as q_{sig} and signal at mapped ME as m_{sig} (fold change relative to genomic average), we model functional conservation as:

$$m_{sig} = f(q_{sig}, g, p) = \sum_{i_1 \in \{0,1\}} \sum_{i_2 \in \{0,1\}} \sum_{i_3 \in \{0,1\}} \alpha_{i_1,i_2,i_3} \cdot \left(q_{sig}^{i_1} \cdot g^{i_2} \cdot p^{i_3} \right),$$

where $g = gkmsim(HE_i, ME_i)$, $p = P_H(ME_i)$, and *f* linearly combines all multiplicative combinations of the three variables (**Figure 6.12**). Predicted mouse regulatory signal $\hat{m}_{sig}$ is computed using 5-fold CV linear regression. $q_{sig} \cdot g \cdot p$, which also has high predictability (**Figure 6.12**) may also be used instead, if m_{sig} is not accessible (e.g., no relevant functional experiments performed in mice).

6.2 Results

gkm-align algorithm finds alignment path of maximum gapped kmer similarity.

Enhancers contain degenerate clusters of TFBS, and enhancer mutagenesis studies have shown that enhancer function is strongly affected by mutations within binding sites and robust to mutations between binding sites[28,29]. This modular architecture more readily tolerates insertions/deletions between TFBS and small structural variations. To exploit this modular structure, *gkm-align* uses a sequence

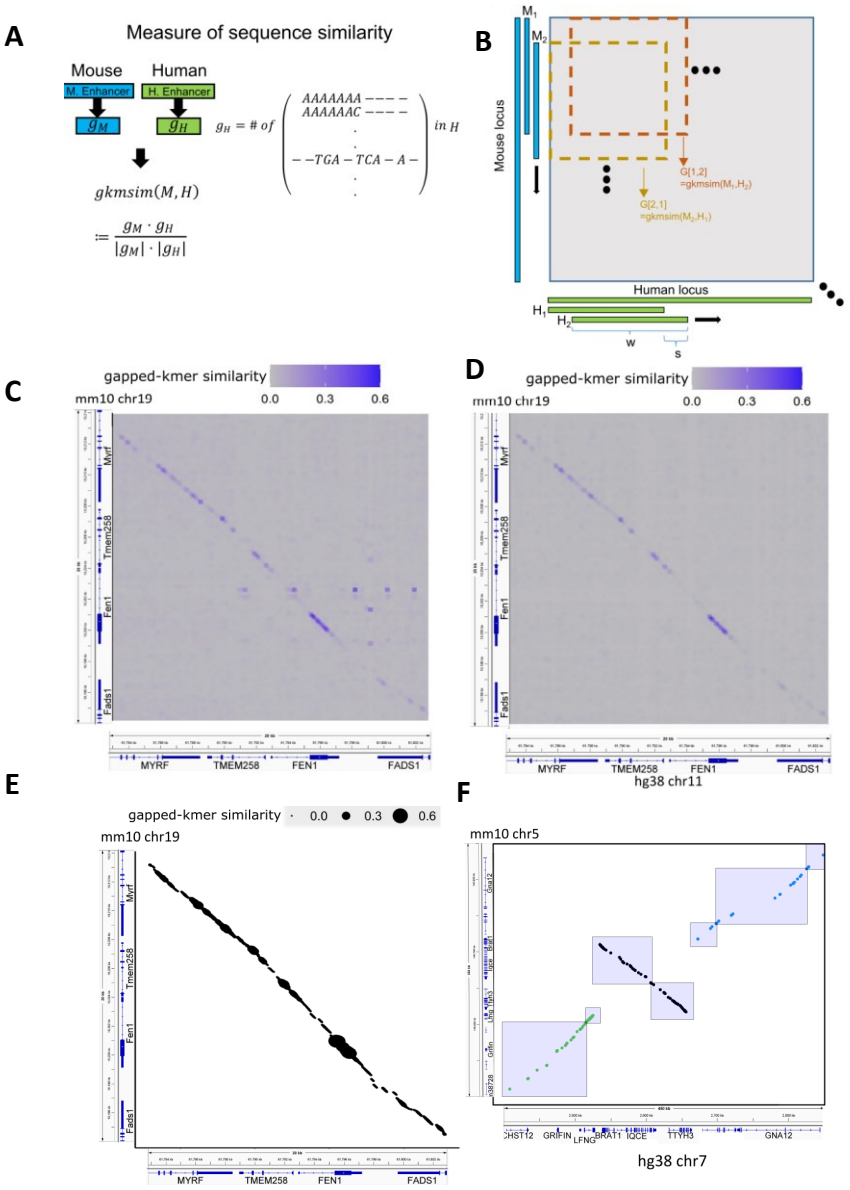

Figure 6.1. Computing pairwise sequence similarity matrix using gapped-kmer composition

A) Sequence similarity between a mouse (M) enhancer and a human (H) enhancer is quantified by their similarity in gapped-kmer compositions (gapped-kmer similarity, or gkm-sim). **B)** Schematic describing computation of pairwise gkm-similarity of all pairs of sliding windows in syntenic genomic loci of the two species. The pairwise similarity values are encoded in a gapped-kmer similarity matrix G. Visualization of gapped-kmer similarity matrix (G) in FADS syntenic locus **C)** without gkm-SVM repeat masking and **D)** with masking. **E)** Identification of colinear series of conserved elements using matrix G. **F)** Schematic describing how whole-genome alignment is performed using the GNA12 inversion locus as an example (dots: short sequence matches. colors: groups of short matches in syntenic blocks; boxes: pairs of human/mouse syntenic loci from which gkm-similarity matrices derive).

similarity metric that compares a pair of sequences (e.g., width of 300 base pairs) by their gapped-kmer composition (**Figure 6.1A; chapter 7** for algorithmic details). Alternative sequence similarity metrics based on kmer-composition have previously been used to detect evolutionarily related sequences[173–175] (typically using 6-mers), but these methods have not been applied to whole genome alignments. We chose to generate whole genome alignments using gapped-kmers because they more accurately model TF binding sites[3,23,30]. Gapped-kmers contain a fixed number of gaps, which represent any nucleotide, and compactly model degenerate positions in TFBS. There exist $\binom{l}{k}4^k$ gapped-kmers with size *l* and *k* non-gapped positions (e.g., N=5,406,720 for *l*, *k*=11, 7), and the gapped-kmer similarity (*gkm-sim*) for a pair of sequences is computed as the cosine similarity of these vectors, each encoding the counts of gapped-kmers in the respective sequence.

To align a pair of human/mouse loci, we first compute a gkm-similarity matrix (*G*) of all pairs of sliding window subsequences of the human and mouse loci (**Figure 6.1B**). The size of this matrix will depend on the locus size; for example, human/mouse loci of

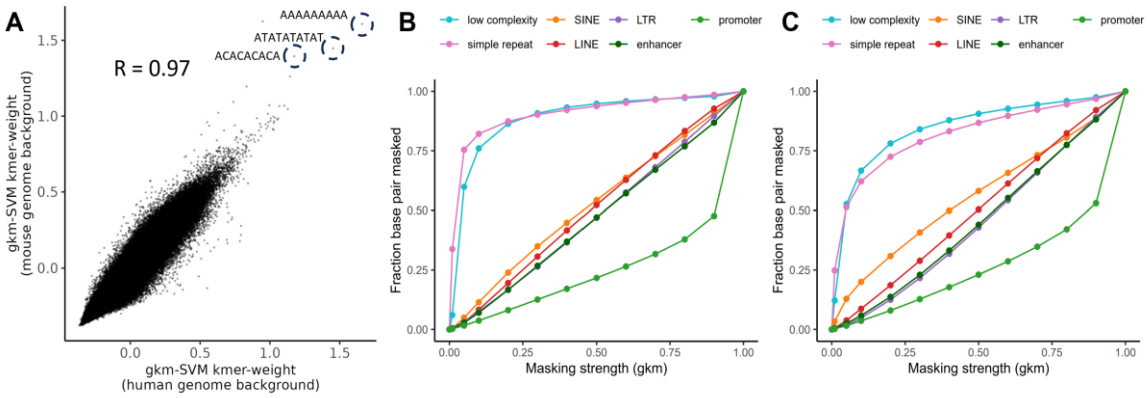

Figure 6.2. Detecting and masking repetitive DNA sequence patterns using gkm-SVM.

A) Comparison of repetitive gkm-SVM vocabularies obtained from human (x-axis) and mouse (y-axis) by kmer weight correlation (R = 0.97). **B)** Fraction of base pairs that are masked by gkm-SVM repeat masking for different classes of repetitive elements at varying masking strengths (method described in more detail in chapter 7) in human and **C)** in mouse.

20 kilobases (e.g., FADS loci) have ~1,000 subsequences of 300 base pair windows sliding by 20 base pairs, and the gkm-similarity matrix (*G*) of dimension 1,000 x 1,000 encodes all pairwise gkm-similarities of the human and mouse subsequences (as shown in **Figure 6.1C**). Exons of matched orthologous genes in human and mouse show highest levels of sequence similarity, but regions of low complexity (e.g., tandem repeats) also show high interspecies sequence similarity due to their prevalence and uniform sequence composition (identifiable as a row of horizontal dots in **Figure 6.1C**). To remove these repetitive sequence matches, we train gkm-SVM to detect and mask ubiquitous across the human and mouse genomes (**Figure 6.1D; Figure 6.2A;** algorithmic detail in **chapter 7.6**). The sequence patterns learned by gkm-SVM, trained on randomly sampled genomic sequences against randomly generated zeroth order DNA sequences (P(A)=…=P(T)), encode background genomic sequences that are prevalent across the genome much higher than expected by chance, such as polyA repeats (**Figure 6.2BC**). We then compute *G* using the masked sequences, to which we apply a variant of Smith-Waterman algorithm to identify an optimal alignment path that encodes how human/mouse loci diverged (**Figure 6.1E**; algorithmic detail in **chapter 7.3**). This method of alignment is extended genome-wide by utilizing orthologous gene annotations[100] and short sequence matches[22,176] (**Figure 6.1F;** algorithmic detail in **chapter 7.4**).

We next demonstrate the *gkm-align* algorithm at the well-studied hemoglobin beta (HBB) locus control region (LCR)[177]. These loci in human and mouse each contains 4-5 enhancers, and the human enhancers have shown to be capable of regulating mouse HBB expression through transgenic mouse experiments[178]. We aligned the HBB LCRs of human and mouse using gkm-align and mapped the five mouse enhancers to human (**Figure 6.3A**). The five mouse enhancers (labeled as ME1, …, ME5) all have

strong DNase I accessibility and EP300 binding in mouse embryonic liver (where HBB is active) and are bound by GATA1[179] in erythroblasts; however only four of the five human loci mapped from these mouse enhancers (labeled as HE1, …, HE5) show strong marks of enhancers. HE5 has weak GATA1 binding in erythroblast and low DNase I accessibility in embryonic liver. Further, inhibiting HE5 with CRISPRi has the weakest effect on downregulating HBE1 expression among the five putative human enhancers mapped from mouse[9,103] (**Figure 6.3B**; **Methods 6.1.2**). It is likely that HE5 may have accumulated mutations leading to loss of regulatory activity. This loss of regulatory activity is also predictable by gapped-kmer sequence similarity metrics. As will be justified further below, **Figure 6.3B** shows that the geometric mean of gapped-kmer sequence similarity and interspecies gkm-SVM enhancer prediction is consistent with the CRISPRi effect at HBB enhancers.

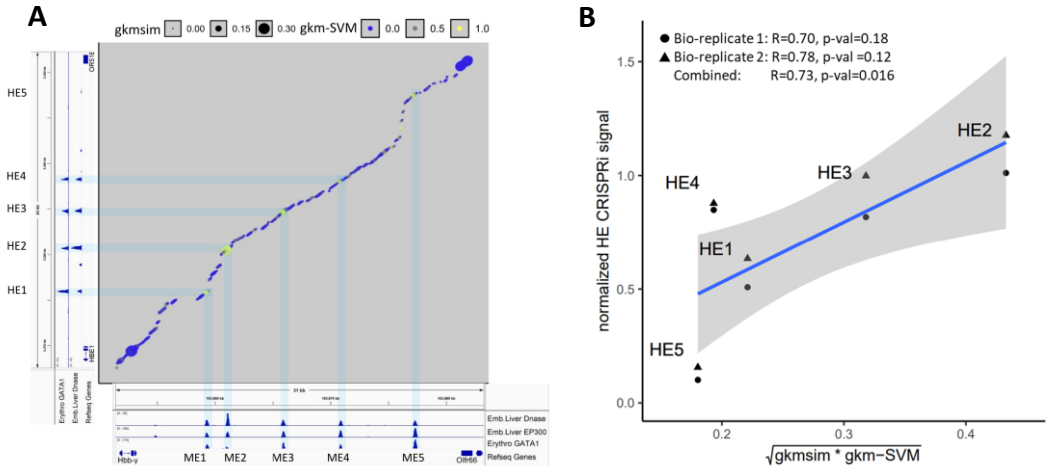

Figure 6.3. gkm-align detects and characterizes conserved enhancers in the HBB LCR

A) Alignment of the HBB Locus Control Region (dot size: gkm-similarity; color: gkm-SVM prediction score at corresponding human locus using gkm-SVM model trained on mouse embryonic liver enhancers. Highlights: CREs); HE: human element; ME: mouse element. **B)** Combining gkm-similarity of (HE, ME) and gkm-SVM score (of HE using mouse embryonic liver enhancer trained model) to predict regulatory activities of human elements (HE) measured by CRISPRi perturbation. Circle: bio-replicate 1, Pearson Corr: 0.7, p-val=0.18; Triangle: bio-replicate 2, Pearson Corr: 0.78; p-val=0.12; Combined: Pearson Corr=0.73, p-val=0.016, linear regression line and 95% confidence interval.

We applied *gkm-align* genome-wide across the 45 cell/tissue pairs and identified many novel conserved enhancers which are not predicted by LASTZ/LiftOver (either predicted to be deleted or which LiftOver maps to inactive regions). Overall, the gkm-align predicted mouse enhancers show clear marks of enhancer activity (**Figure 6.4A**): strong DNase/H3K27ac/H3K4me1 signals and weaker H3K4me3 (averaged across 9 cell/tissues for which the mouse histone ChIP-seq data are available; the cell/tissue identities are shown in **Supplementary Table 1**). For all the 45 cell/tissue pairs, gkm-align mapped a higher number of human enhancers to mouse enhancers than LASTZ/LiftOver (**Figure 6.5**), with the increase in enhancer mappability ranging from 1% (embryonic limb) to 22% (hematopoietic multipotent progenitor cells) (c=0 in **Figure 6.8A**). For the cell/tissue pair of human hippocampus astrocyte and mouse Müller cells (both of which are glial cells), *gkm-align* successfully mapped 791 human enhancers to

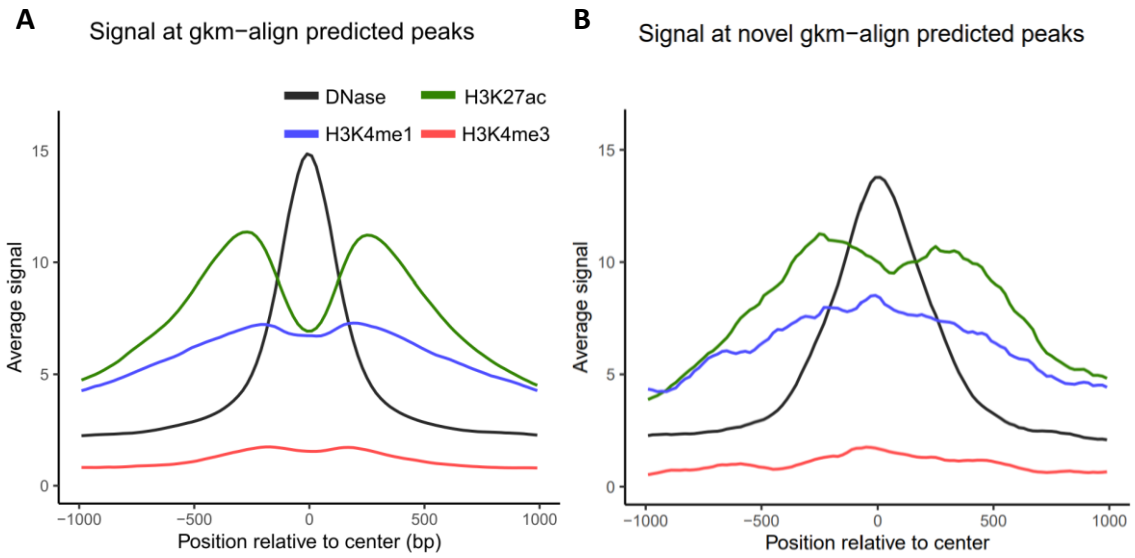

Figure 6.4. gkm-align predicted mouse enhancers show clear marks of enhancer activity

A) Average DNase accessibility and H3K27ac/H3K4me1/H3K4me3 histone ChIP-seq signals at mouse loci mapped from human enhancers (aggregated across 9 distinct cell/tissues) using gkm-align. Signals are normalized as fold-change from genomic average. **B)** same but constrained to novel conserved mouse enhancers mappable uniquely by gkm-align (c=0).

mouse enhancers, which are incorrectly mapped by LASTZ/LiftOver (either deleted or map to inactive mouse regions). These novel conserved enhancers show clear markers of enhancer activity (**Figure 6.4B**). Conversely, only 222 human enhancers were correctly mappable uniquely by LASTZ/LiftOver but incorrectly mappable by *gkm-align*. 8,559 human glial enhancers were mapped to mouse enhancers by both methods. Together across the 45 cell/tissues, gkm-align identified 6,591 novel conserved enhancers. This greatly increases the number of human enhancers which can be functionally tested for disease relevance in mouse models.

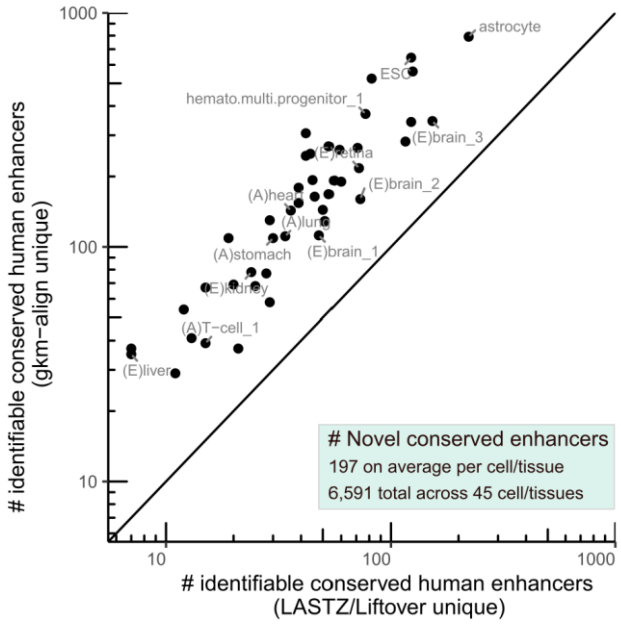

Figure 6.5. gkm-align discovers thousands of novel conserved enhancers.

Number of human-mouse conserved enhancers that are identifiable uniquely by LASTZ/LiftOver (x-axis) and gkm-align (y-axis) for each of the 45 cell/tissue pairs. Gkm-align identifies conserved enhancers missed by LASTZ/LiftOver in all tissues. Relative to LASTZ/LiftOver, gkm-align discovers 197 novel enhancers on average per cell/tissue and total 6,591 novel enhancers across all 45 cell/tissues.

gkm-align identifies additional novel conserved enhancers when combined with cell-specific vocabulary.

Although *gkm-align* outperforms LASTZ/LiftOver, the sequence similarity metric does not explicitly make use of cell/tissue specific regulatory vocabulary. Gkm-SVM enhancer regulatory vocabularies, encoding TFBS motifs, are well conserved between human and mouse (**Figure 4.9B, Figure 5.1A**), and they can be incorporated into gkm-align both to improve discovery of conserved enhancers and to quantify their predicted functional conservation. This additional information leads to an expanded catalog of human enhancers testable through mouse models, ranked by likelihood of conserved regulatory roles.

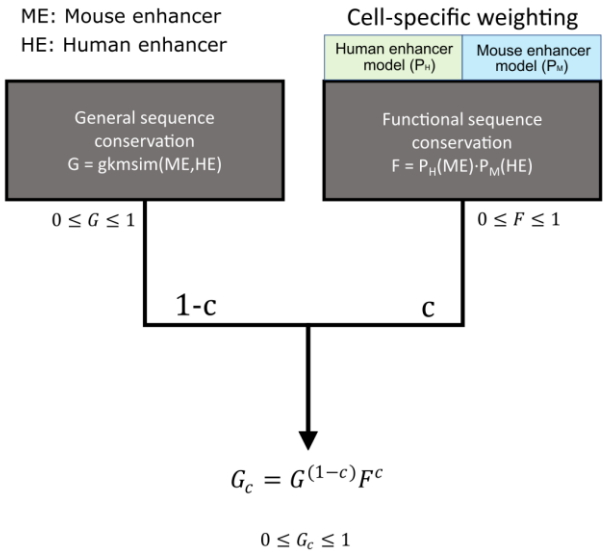

Figure 6.6. Cell-specific genome-alignment using gkm-SVM enhancer models

Schematic describing how cell-specific gkm-SVM enhancer prediction model is incorporated into gkm-align for cell-specific weighted alignment.

We incorporate cell-specific gkm-SVM regulatory vocabularies into gkm-align following a simple and intuitive model: if a pair of human and mouse enhancers, denoted as HE and ME, are orthologous, then they should have similar DNA compositions (i.e.,

general sequence conservation), and should both contain conserved TFBS motifs relevant to the shared cellular context (i.e., functional sequence conservation) (**Figure 6.6**). General sequence conservation (*G*) is quantified using gkm-similarity, as previously described. Functional sequence conservation (*F*) is computed using interspecies gkm-SVM prediction scores, which we normalize to vary between 0 and 1 for interpretability (algorithmic detail in **chapter 7.5**). Denoting $P_H(ME)$ as normalized prediction score of a mouse element by a gkm-SVM model trained on human enhancers, we can interpret $P_H(ME)$ as the probability that *ME* can function as an enhancer in the orthologous human cellular context. $P_M(HE)$ is defined similarly (M: mouse; HE: human enhancer). Then, functional sequence conservation, computed as $F = P_H(ME) \cdot P_M(HE)$, can be interpreted as the probability that *ME* and *HE* can both function as enhancers interchangeably in human and mouse cellular contexts. For cell-specific weighted alignment, we combine the two measures of enhancer conservation into $G_c = G^{1-c} \cdot F^c \ (0 \leq c \leq 1)$, which adjusts the alignment path toward human/mouse element pairs with both similar sequence composition (*G*) and functional similarity in common cellular context (*F*).

Cell-specific weighted alignment by *gkm-align* identifies the highest number of conserved enhancers when it is combined with gkm-SVM enhancer prediction model trained on enhancers of relevant cell/tissue type. For example, LASTZ/LiftOver and gkm-align each identify 1,325 and 1,529 conserved human B-cell enhancers, but if gkm-align is combined with B-cell trained gkm-SVM enhancer prediction models, the number of identifiable conserved enhancers increases up to 1,818 at cell-specific enhancer model weighting parameter (*c*=0.8) (a 37% increase from LASTZ/LiftOver) (**Figure 6.7A**). The identification rate also increases when *gkm-align* is combined with gkm-SVM models of similar cell-types (with overlapping TFs), such as myeloid progenitor cells and thymus,

each with peaks at 1,663 (*c*=0.5) and 1,605 (*c*=0.4) conserved enhancers. Similarly for colon enhancers, LASTZ/LiftOver, unweighted gkm-align (*c*=0), and cell-specific gkm-align (*c*=0.75) each identifies 1,125, 1,221, and 1,352 conserved enhancers, which corresponds to a 7.9% and 20% increase over LASTZ/LiftOver for unweighted (*c*=0) and weighted (*c*=0.75) gkm-align respectively (**Figure 6.7B**). Cell-specific weighting using enhancer-trained gkm-SVM models improves the identification rate of conserved enhancers for all pairs of 45 cell/tissues (**Figure 6.8A**). At *c*=0.9, we observe up to an 80% increase in conserved enhancer discovery over LASTZ/LiftOver for monocytes, with 16 cell/tissues with greater than a 20% increase for *c*=0.7 and c=0.8. A subset of cell/tissues exhibited limited improvement through gkm-SVM weighting (e.g., brain), but their identification rates remained higher than both unweighted gkm-align and LASTZ/LiftOver at *c*=0.5. Across the 45 cell/tissues, weighted cell-specific gkm-align discovers several hundred novel enhancers in every tissue and 23,660 total novel

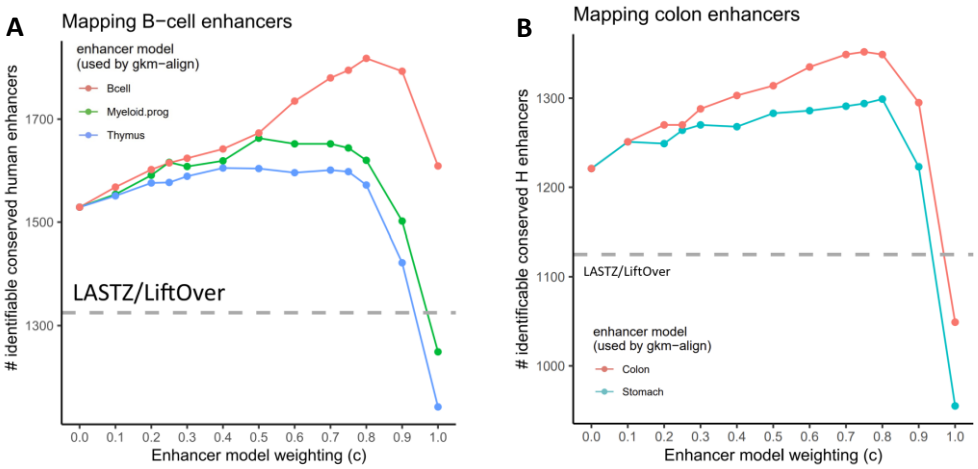

Figure 6.7. Enhancing discovery of conserved enhancers by incorporating gkm-SVM enhancer vocabularies.

A) Number of B-cell human enhancers mappable to mouse B-cell enhancers using LASTZ/LiftOver (grey dashed line) and gkm-align weighted by gkm-SVM enhancer models trained on B-cell (red), myeloid progenitor cell (green), and thymus enhancers (blue) with varying weights. **B)** Number of human colon enhancers mappable to mouse colon enhancers using LASTZ/LiftOver (grey dashed line) and gkm-align weighted by gkm-SVM enhancer models trained on colon (red) and stomach (blue) with varying enhancer model weights.

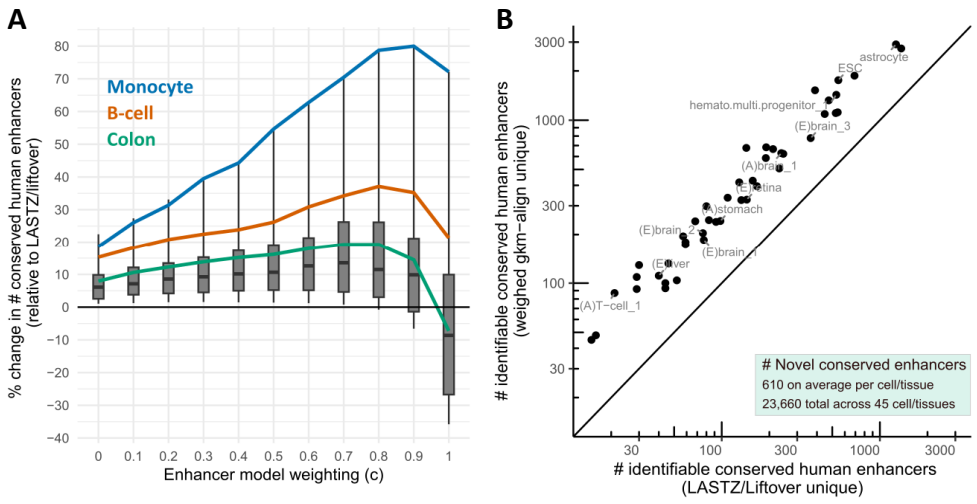

Figure 6.8. gkm-align detects higher numbers of conserved enhancers when combined with

A) Percent change in the number of identifiable human/mouse enhancer pairs by gkm-align, relative to LASTZ/LiftOver, across 45 cell-tissue pairs using cell-type matched gkm-SVM model weighting. Line plots for a subset of cell/tissues. **B)** Number of human-mouse conserved enhancers that are identifiable uniquely by LASTZ/LiftOver (x-axis) and gkm-SVM weighted cell-specific gkm-align (y-axis) for each of the 45 cell/tissue pairs. Relative to LASTZ/LiftOver, weighted gkm-align discovers 610 novel enhancers on average per cell/tissue and total 23,600 novel enhancers across all 45 cell/tissues.

conserved enhancers across all 45 cell/tissues (**Figure 6.8B**). Despite the increased performance by gkm-align with or without gkm-SVM cell-specific weighting, the cell-specific pattern of enhancer conservation we showed in **Figure 5.2A** is maintained (**Figure 6.9**). Lastly, it should be noted that, although cell-specific alignment discovers a higher number of conserved enhancers compared to unweighted *gkm-align*, there are also handful of enhancers identifiable by unweighted gkm-alignment but missed by cell-specific gkm-alignment. These tend to be conserved enhancers with high general sequence similarities but with more degenerate TFBS. For this reason, both generic and cell-specific *gkm-align* should be used in parallel to maximize the span of identifiable conserved enhancers, and we provide genome-wide enhancer mappings for both methods.

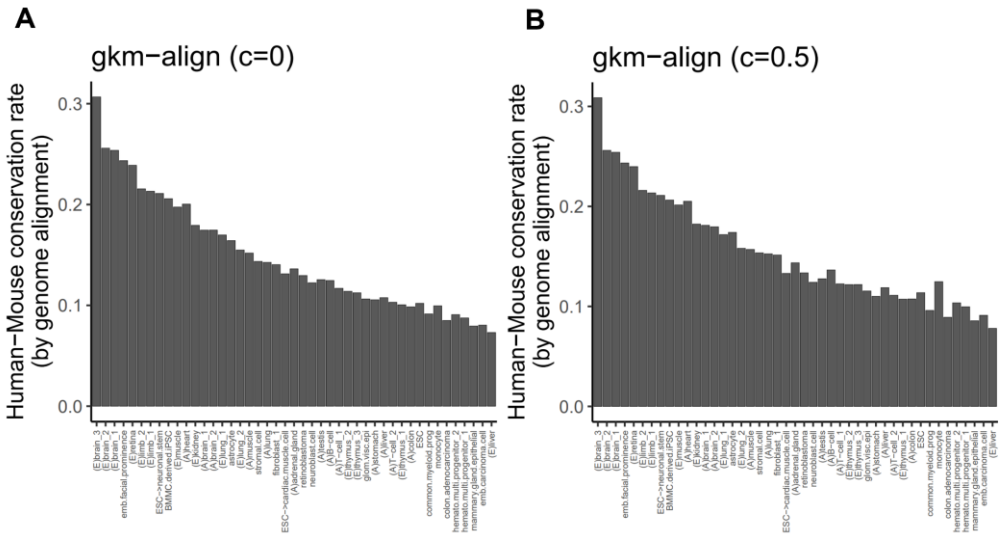

Figure 6.9. Conservation analysis using gkm-align is consistent with the analysis using LASTZ/liftover in Fig 5.2A

Same as the enhancer conservation rates shown in Fig 5.2A but using gkm-align with cell-specific **A)** no weighting (c=0) and **B)** weighting (c=0.5).

We reproduced the above results using an independent dataset[93], where enhancers in diverse non-human mammals in four tissues (brain, muscle, liver, testis) were identified using H3K27ac and H3K4me1 histone marks. The four tissues are also included among our list of 45 cell/tissues, and we evaluated gkm-align by mapping our enhancers (distal cell-specific DHS) in these tissues to the histone-defined enhancers. Consistent with **Figure 6.8**, Gkm-align maps higher numbers of conserved enhancers than LASTZ/LiftOver across four tissues, where gkm-align's performance maximally improves when tissue-matched gkm-SVM enhancer models are incorporated (**Figure 6.10**). Using the same dataset, we also show that gkm-align also maps higher numbers of conserved human enhancers in rhesus macaque for all the four tissues (**Figure 6.11**). For this analysis, we used gkm-SVM models trained on our mouse enhancer sets to align human and macaque, which highlights the applicability of gkm-SVM models in

different species due to conservation of enhancer vocabularies (**Figure 4.9B**, **Figure 5.1A**).

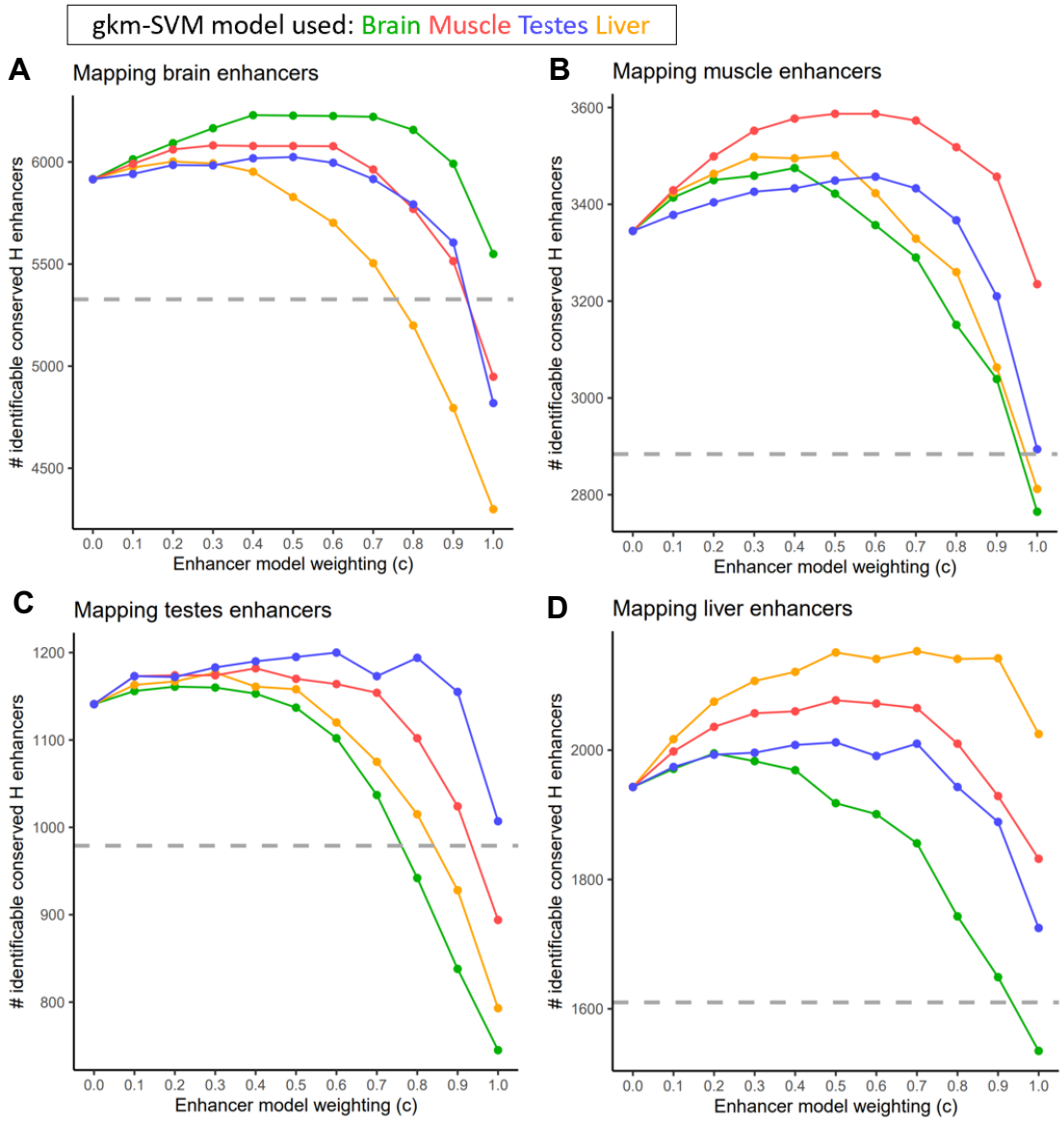

Figure 6.10. Evaluation of gkm-align predictions on the independent mouse enhancer sets of Roller 2021.

Number of human **A)** brain, **B)** muscle, **C)** testes, and **D)** liver enhancers (distal cell-specific DHS) mappable to mouse enhancers (H3K27ac/H3K4me1) in matched tissues using LASTZ/LiftOver (grey dashed line) and gkm-align weighted by gkm-SVM enhancer models trained various tissues. various tissues.

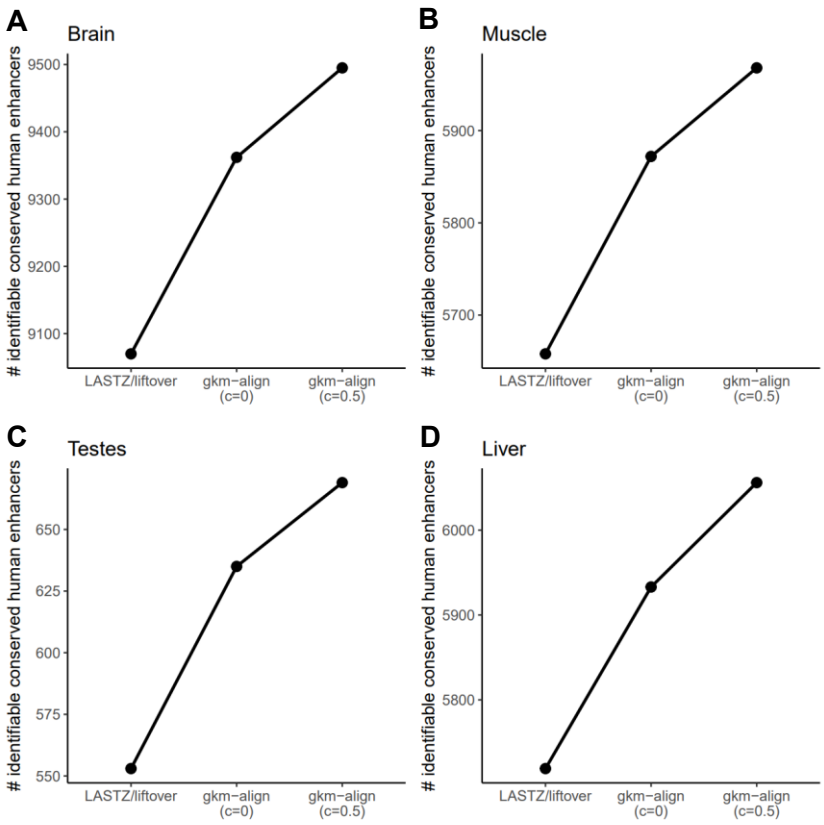

Figure 6.11. Evaluation of gkm-align predictions on the independent macaque enhancer sets of Roller 2021.

Number of human **A)** brain, **B)** muscle, **C)** testes, and **D)** liver enhancers (distal cell-specific DHS) mappable to macaque enhancers (H3K27ac/H3K4me1) in matched tissues using LASTZ/liftOver and unweighted gkm-align, and gkm-align weighted by gkm-SVM models trained on enhancers in matched tissues.

Cell-specific information from gkm-SVM enhancer prediction models can also be combined with the gapped-kmer based sequence similarity metric (*gkm-sim*) to quantify functional conservation of orthologous enhancer pairs discovered through *gkm-align*. The strength (DHS signal) of mouse loci mapped from human enhancers tends to correlate with strength of the human enhancers, but a fraction of the identified orthologous pairs lacks such conservation of activity due to sequence divergence (**Figure 5.1B**). To rank predictions, we explored different ways to predict functional conservation between orthologous human and mouse enhancer pairs identified by *gkm-align* (**Figure 6.12A**). We observed that the DNase signal of a query human sequence

(qDNase) was correlated with the DNase signal of the mapped mouse ortholog (mDNase) with median correlation of 0.29 (min: 0.13; max: 0.46), and the product of qDNase and gkm-similarity between query human enhancer (HE) and the mapped mouse ortholog (ME), lead to increased median correlation of 0.41 (min: 0.23; max: 0.60) (**Figure 6.12B**). When this product is further multiplied by enhancer prediction of ME by human trained gkm-SVM ($0 \leq P_H(ME) \leq 1$), median correlation further improves to 0.46 (min: 0.25; max: 0.61). This triple product has a high value for a mapped mouse ortholog if it has similar gapped-kmer composition, a strong human enhancer ortholog, and contains conserved TFBS motifs. Training a regression model with these combinatorial features leads to median correlation of 0.55 (min: 0.32; max: 0.68) (**Figure 6.12B**; **Methods**). All of our human/mouse conserved enhancer mappings are ranked with these gapped-kmer based conservation scores, which we believe will facilitate downstream experimental testing by providing confidence ranking scores for functional conservation in mice.

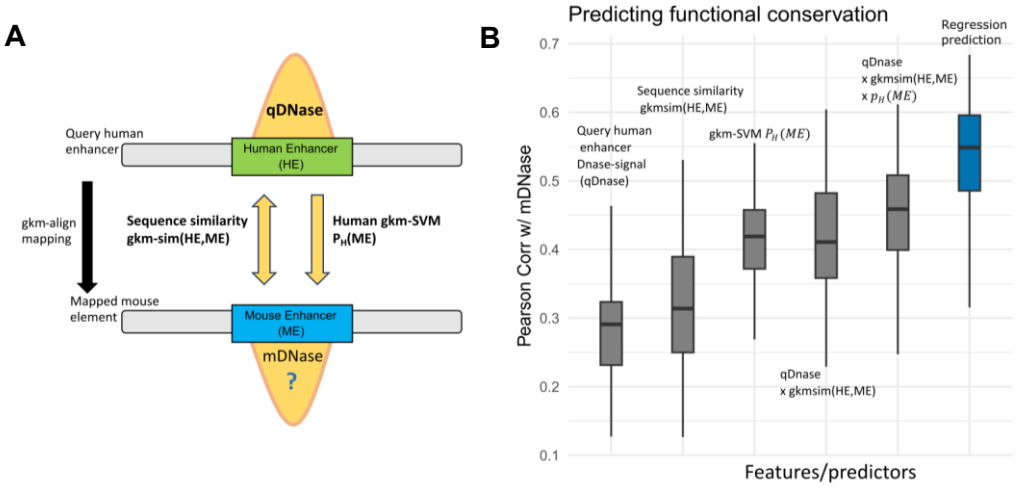

Figure 6.12. gkm-align identifies more novel conserved enhancers and robustly predicts functional conservation when combined with cell-specific information.

A) Schematic describing regression model for ranking enhancer mapping to mouse (mDNase) in terms of human enhancer features (qDNase = DNase-signal at query human enhancer; gkm-sim = gapped-kmer sequence similarity between human and mouse element; PH(M) = human gkm-SVM score of mapped mouse element). **B)** Predicting DNase-seq signal at mouse loci mapped from human enhancers (mDNase) using combinations of features described in Figure 4E across the 45 human/mouse cell/tissue pairs.

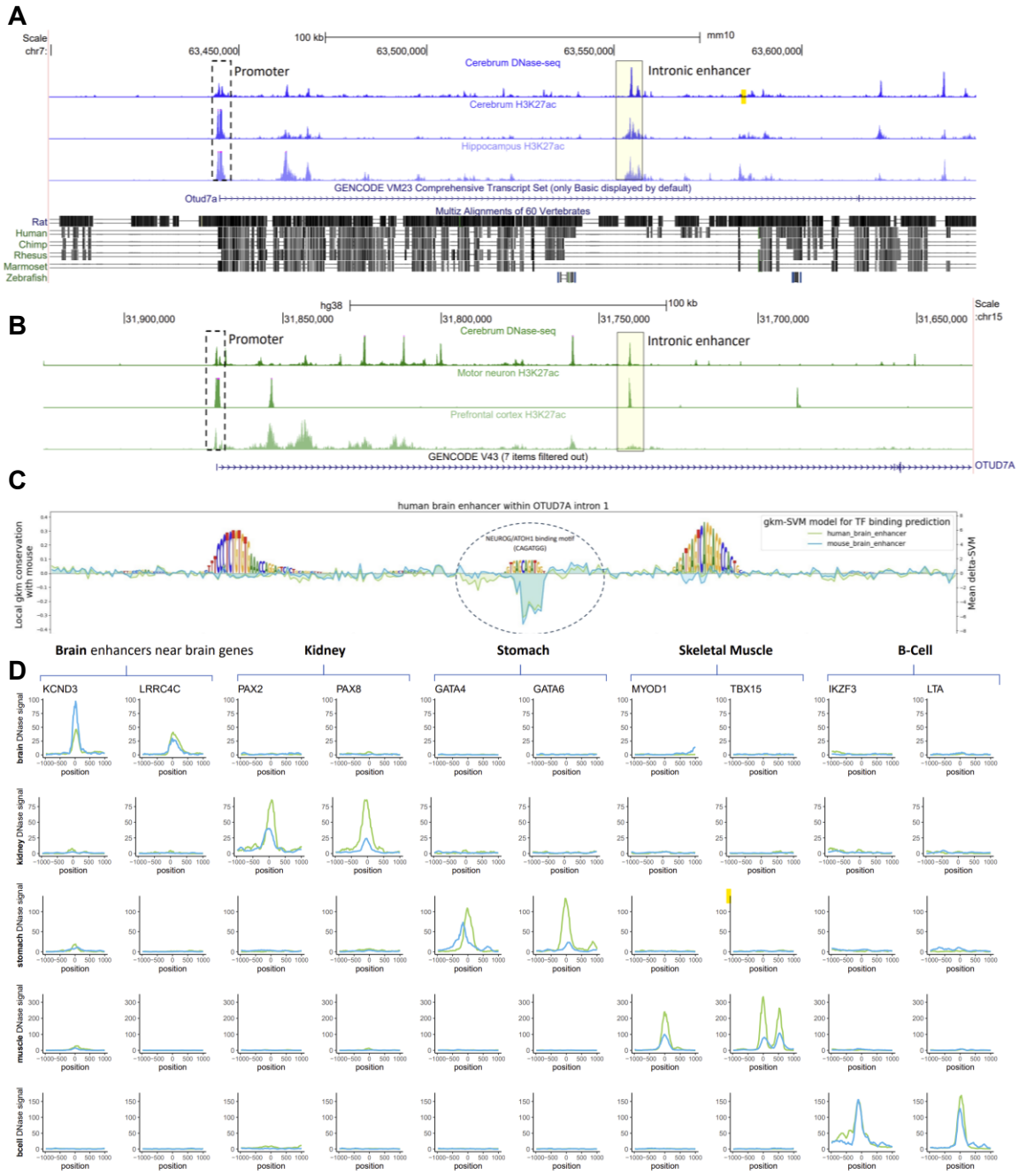

Figure 6.13. Examples of novel enhancers from expanded catalogue of human/mouse orthologous enhancers.

A) Genome browser visualization of mouse and **B)** human OTUD7A loci. Yellow boxes indicate the identified conserved enhancers. **C)** Visualization of conserved binding sites in human OTUD7A intronic enhancer by sequence conservation with the orthologous mouse enhancer. Logo height represents local gapped-kmer sequence conservation score with mouse, and line plots indicate TF-binding prediction by delta-SVM models trained on human or mouse brains enhancers. **D)** Visualization of DNase accessibility of orthologous enhancers from the five distinct cell/tissues (Green: human, Blue: mouse; position relative to DHS center of human/mouse orthologous enhancers; signal: fold change from genomic average).

Many of the novel enhancers identified by *gkm-align* are supported by additional evidence of conserved function. For example, gkm-align predicts a conserved enhancer in OTU Deubiquitinase 7A (OTUD7A) which is highly expressed in both human and mouse brains[32] and is associated with a wide range of neurological diseases as such schizophrenia and epilepsy[180]. OTUD7A knockout leads to morphological deformation of cortical neurons and frequent seizure-like events in mice[181,182]. We identified orthologous pairs of putative OTUD7A enhancers in intron 1 of human and mouse OTUD7A (**Figure 6.13A-B**; enhancers: yellow-highlighted; hg38/chr15:31740273-31740573; mm10/chr7:63554547-63554847). Both human and mouse elements exhibit strong DNase-I accessibility and H3K27ac histone modification across biological samples related to the nervous system (**Figure 6.13A-B).** The two enhancers appear to have three clusters of conserved DNA base pairs with high local conservation in gapped-kmer composition (local conservation rate represented as the logo heights in **Figure 6.13C**), and one of the clusters located at the centers of the human and mouse enhancers contains a NEUROG/ATOH1 binding motif (GCAGATGG), which is identified among the top brain enhancer kmer weights for both human and mouse as shown in **Figure 4.7**. This part of the enhancer has the largest delta-SVM[3] score for gkm-SVM models trained on both human and mouse brains (visualized using shaded line plots in Fig 5C), indicating that it is a core TF binding site conserved between human and mouse. Despite the clear conserved biochemical signatures and binding motif, this enhancer is predicted to be deleted in mouse by LASTZ/LiftOver.

To show further examples of the top conserved enhancers in **Figure 6.13D,** we ranked enhancers that have the strongest combined (i) DNase I accessibility, (ii) gkm-similarity, and (iii) interspecies gkm-SVM prediction, using the regression score described in **Figure 6.12**. This combined regression score increases the likelihood of

functional conservation as shown in **Figure 6.3B** for HBB CRISPRi. We ranked enhancers collected from five diverse human and mouse cell/tissues (brain, kidney, stomach, muscle, B-cell) from among the top 1% conserved enhancers with highest regression score. Among these top orthologous enhancers, we selected a subset of enhancers in the vicinity of orthologous genes with cell/tissue specific expression[32] (**Figure 6.13D**). These genes include KCND3 (brain; voltage-gated potassium channel subunit), PAX2 (kidney; TF associated with renal malformation[183]), GATA6 (stomach; definite endoderm TF[8]), MYOD1(muscle; TF associated with myopathy[184]), and IKZF3 (B-cell; TF mutated in leukemia[185,186]). For each of the 45 cell/tissue pairs, we generated a table of ranked orthologous human-to-mouse enhancer pairs. In addition to providing an expanded catalog of conserved distal enhancer elements, the ranking can be used to prioritize elements for functional characterization. We uploaded our catalogue of conserved enhancers to beerlab.org/gkmalign/.

6.3 Discussion

We developed a gapped-kmer based novel alignment algorithm to detect conserved enhancers, *gkm-align. Gkm-align* maps orthologous enhancers by finding alignment paths of maximal gapped-kmer composition at the resolution of sliding ~300 base pair windows. We used a whole genome alignment strategy and present a new set of conserved enhancer predictions for human and mouse. We evaluated these predictions on 45 pairs of matched tissues using ENCODE data and show that *gkm-align* detects thousands of conserved enhancers missed by conventional alignment methods. We further extend these predictions by combining tissue specific TF information, which predicts an additional 500 enhancers per tissue on average, and up to an 80% increase in some tissues. While our analysis confirms that mapping orthologous enhancers

between distant mammals is an inherently difficult problem due to rapid enhancer evolution, we show that we detect conserved enhancers of biomedical significance missed by LiftOver/LASTZ, including an intronic enhancer of OTUD7A which is associated with epilepsy in humans and when knocked out reduces dendritic density and promotes seizures in mouse. Many multiple alignment algorithms[187–189] build upon pairwise alignment outputs from LASTZ, and we expect that the improvement of pairwise genome alignment by *gkm-align* will lead to improved annotations of conserved cis-regulatory elements in diverse mammalian genomes that are not functionally characterized as deeply as human and mouse. Despite the algorithmic improvement for mapping orthologous enhancers, our analysis confirms the overall weak enhancer conservation relative to promoters and that enhancers have surprisingly variable conservation rate across cells/tissues. Lastly, we provide an expanded catalogue of orthologous human-mouse enhancers, each annotated with predictive gapped-kmer based functional conservation scores. We expect that this expanded and quantitatively ranked catalogue of conserved enhancers will facilitate discovery and functional characterization by prioritizing enhancers for testing in model animals.

Chapter 7

Algorithmic details of gkm-align

In this chapter, I provide algorithmic details of gkm-align[8]. I first introduced the gkm-align algorithm in the previous chapter, which points to sections in this chapter for more detailed descriptions of the algorithm. This chapter provides supplementary information to the previous chapter. The gkm-align algorithm was motivated by previous successes of machine learning models using gapped-kmers in effectively learning sequence features of enhancers and discover their regulatory vocabularies[3,23,30]. The gkm-align algorithm utilizes gapped-kmers as sequence feature, and identifies conserved enhancers by finding optimal alignment paths of maximum gapped-kmer similarity.

7.1 Algorithm overview

Given a human-mouse syntenic locus, gkm-align identifies an alignment path spanning the human and mouse sequences (S_H, S_M). The alignment is done at the resolution of sliding windows (**Figure 7.1A**; e.g., window width, $\mathrm{w} = 300$; slide step size, $s = 20$). We first compute pairwise sequence similarities (using the gapped-kmer kernel, i.e. the dot product of gapped-kmer counts) between all pairs of human and mouse sliding windows, and encode the pairwise sequence similarity values in the matrix G (**Figure 7.1B**; details in **chapter 7.2**). $G[i,j]$ encodes the gapped-kmer similarity between the i^{th} and j^{th} sliding windows of S_H and S_M (denoted as $S_{H,i}$ and $S_{M,j}$). Repetitive portions of the human and mouse sequences are masked prior to computing G (**chapter 7.5**). The resulting matrix G is then used to identify the optimal alignment path with maximal gapped-kmer similarity. (**Figure 7.1.C**; details in **chapter 7.3**). Gkm-align is extended genome-wide by

[8] This chapter is adapted from a supplementary document of the gkm-align manuscript[10]

defining human-mouse syntenic blocks using the ENCODE human-mouse orthologous gene coordinates **Fig 7.1C**; details in **chapter 7.3**). Human-mouse short sequence matches are further used to obtain finer syntenic blocks, leading to higher computational

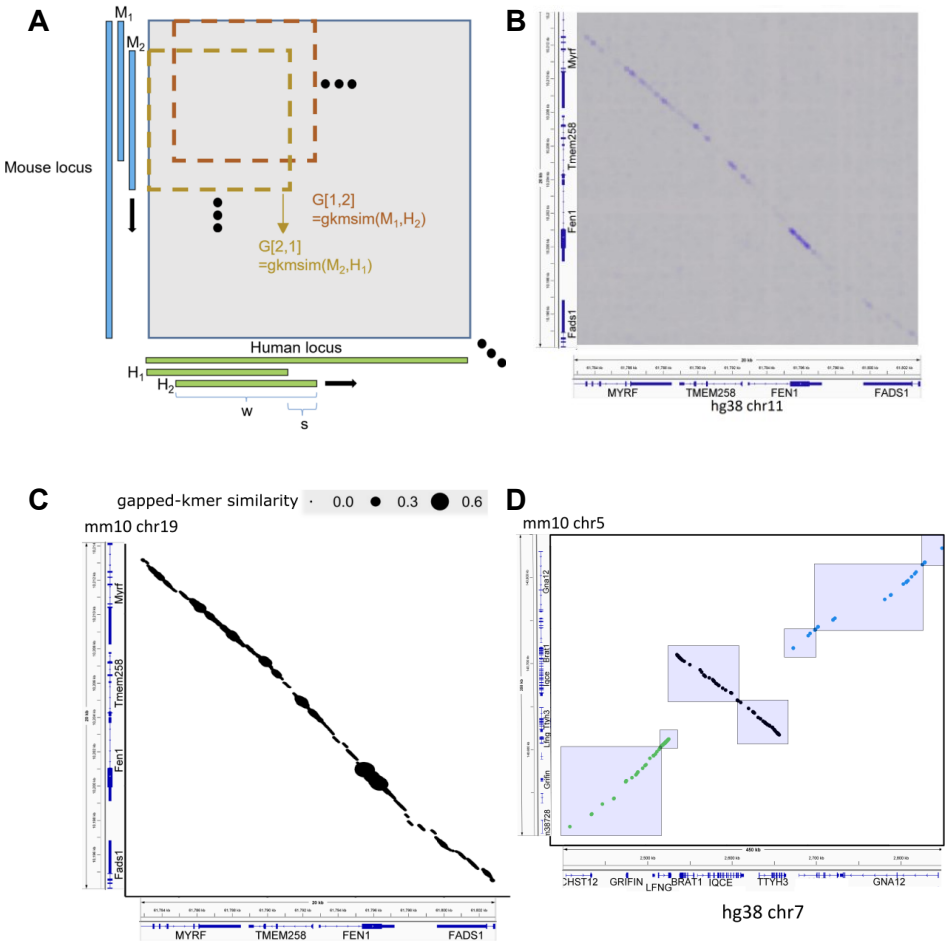

Figure 7.1. Graphical overview of gkm-align

A) Sequence similarity between a mouse (M) enhancer and a human (H) enhancer is quantified by their similarity in gapped-kmer compositions (gapped-kmer similarity, or gkm-sim). The schematic describes computation of pairwise gkm-similarity of all pairs of sliding windows in syntenic genomic loci of the two species. The pairwise similarity. **B)** gapped-kmer similarity matrix (G) in FADS syntenic locus with gkm-SVM repeat masking. **C)** Identification of colinear series of conserved elements using matrix G. G) Alignment of the HBB Locus Control Region (dot size: gkm-similarity; color: gkm-SVM prediction score at corresponding human locus using gkm-SVM model trained on mouse embryonic liver enhancers. Highlights: CREs); HE: human element; ME: mouse element. **D)** Schematic describing how whole-genome alignment is performed using the GNA12 inversion locus as an example (dots: short sequence matches. colors: groups of short matches in syntenic blocks; boxes: pairs of human/mouse syntenic loci from which gkm-similarity matrices are derived).

efficiency of whole-genome alignment and resolution of small structural variations (**Fig 7.1D**; details in **chapter 7.4**). Lastly, we can further improve conserved enhancer discovery of gkm-align by incorporating gkm-SVM cell-specific enhancer regulatory vocabularies (**chapter 7.5**).

7.2 Optimal computation of gapped-kmer matrix G constructed from overlapping sliding windows.

At the core of the gkm-align algorithm is computation of sequence similarity between a pair of DNA sequences (300 base pairs long by default) by their gapped-kmer sequence composition: $G(S_1, S_2)$. Using Eq. 4.2 from **chapter 4**, gkm-similarity of the i^{th} and j^{th} sliding windows of S_H and S_M, denoted as $S_{H,i}$ and $S_{M,j}$ (**Figure 7.1.A**) is computed as the cosine similarity of their gapped-kmer compositions:

$$G_{S_H,S_M}[i,j] = \frac{g(S_{H,i}) \cdot g(S_{M,j})}{\|g(S_{H,i})\| \cdot \|g(S_{M,j})\|}$$

(eq. 7.1)

Efficient computation matrix G requires efficient computation of $g(S_1) \cdot g(S_2)$ for any DNA sequences S_1 and S_2. The first step of reducing gkm-align's time-complexity is using Eq. 4.5, derived in **chapter 4**:

$$g(S_1) \cdot g(S_2) = \sum_{i=1}^{|S_1|-l+1} \sum_{j=1}^{|S_2|-l+1} \begin{cases} \binom{l-m_{ij}}{k}, l-k \geq m_{ij} \\ 0 \quad , l-k < m_{ij} \end{cases}$$

, where m_{ij} denotes the number of mismatched base pairs between $g(S_1(i,l)) \cdot g(S_2(j,l))$. Gkm-SVM[23] efficiently computes the mismatch profile using a kmer-tree structure. Gkm-align computes the mismatch profile using SIMD (Single instruction, multiple data) parallel

computation, which is more efficient for its algorithmic structure (**Figure 7.2**). For a pair of human and mouse kmer sequences, gkm-align uses SIMD to compare the two kmers simultaneously for all base pair positions, and this optimization technique allows gkm-align's time complexity to be nearly independent of l (kmer size).

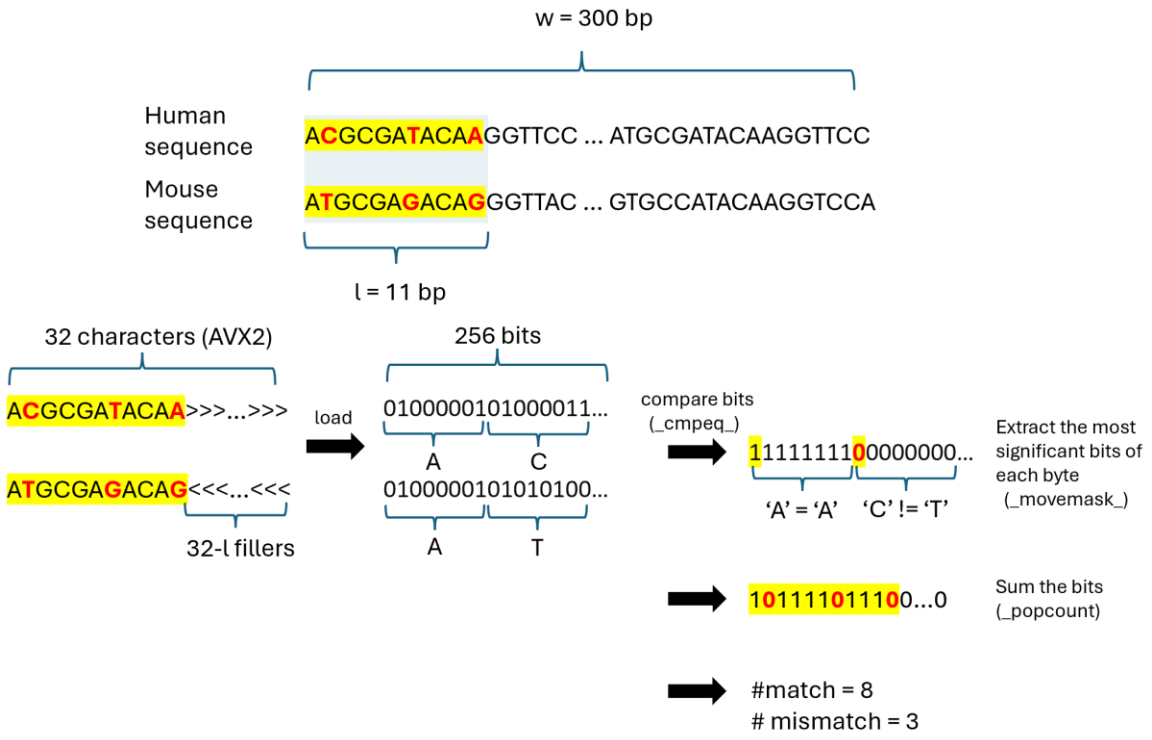

Figure 7.2. Efficient computation of kmer nucleotide mismatches using SIMD (single instruction multiple data) parallel computation.

G encodes pairwise sequence similarities of sliding windows from human and mouse syntenic loci, and computation of G can be significantly optimized by minimizing redundant computations in Eq. 6.1 coming from overlaps with neighboring sliding windows. Recall that $S_{H,i}$ and $S_{M,j}$ each represent i^{th} and j^{th} sliding windows of S_H and S_M. The algorithm requires that w (sliding window width) is divisible by s (slide step size). Then, $S_{H,i}$ and $S_{M,j}$ can each be represented as a concatenated ($\sim$) series of contiguous s-mers, repeating w/s times (**Figure 7.3**). That is:

$$S_{H,i} = S_H(1 + (i-1) \cdot s, s) \sim S_H(1 + i \cdot s, s) \sim S_H(1 + (i+1) \cdot s, s) \sim$$

$$\ldots \sim S_H(1 + (i + w/s - 2) \cdot\ s, s)$$

$$S_{M,j}\ =\ S_M(1 + (j - 1) \cdot\ s, s) \sim S_M(1 + j \cdot\ s, s) \sim S_M(1 + (j + 1) \cdot\ s, s) \sim$$

$$\ldots \sim S_M(1 + (j + w/s - 2) \cdot\ s, s)$$

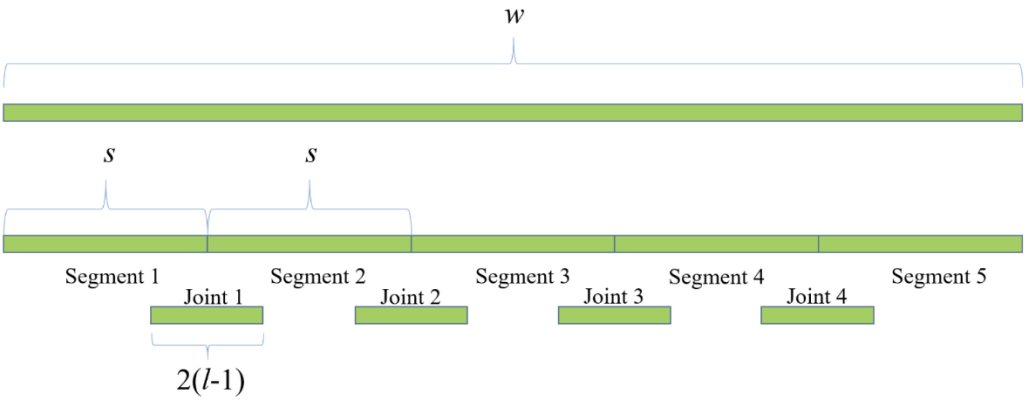

Figure 7.3. Optimizing computation of *G* deriving from overlapping sliding genomic windows

Before deriving gkm-composition of the above two representations, let's consider the following observation. Suppose $S = S_A \sim S_B$, where $w_A\ = |S_A|, w_B = |S_B|, s \geq\ l$. Then, $g(S) \geq\ g(S_A)\ +\ g(S_B)$ since S contains not only subsequences of S_A and S_B but also subsequences that are newly introduced by joining S_A and S_B together. Let's call such subsequence of S as $joint(S_A, S_B) = S(w_A - l + 1, 2(l - 1))$. Then,

$$g(S) = g(S_A \sim S_B) = g(S_A)\ +\ g(S_B)\ +\ g(joint(S_A, S_B))$$

Using the same idea, then

$$g(S_{H,i})\ =\ g(S_H(1 + (i - 1) \cdot\ s, s) \sim S_H(1 +\ i \cdot\ s, s) \sim \ldots \sim S_H(1 + (i + w/s - 2) \cdot\ s, s)$$

$$= \sum_{p=1}^{w/s} g(S_H(1 + (i + p - 2) \cdot\ s, s))$$

$$+ \sum_{q=1}^{w/s-1} g(joint(S_H(1 + (i + q - 2) \cdot s, s), S_H(1 + (i + q - 1) \cdot s, s)))$$

, where the first series computes the total gapped-kmer composition at individual segments being concatenated and the second series computes the total composition at the joints. Naming the terms in the first series as segments and those in the second series as joints for simplicity, we get the following representations for the gapped-kmer compositions of the human and mouse sliding windows.

$$g(S_{H,i}) = \sum_{p=1}^{w/s} (p^{th}\ segment\ of\ S_{H,i}) + \sum_{q=1}^{w/s-1} q^{th}\ joint\ of\ S_{H,i}$$

(Eq. 7.2)

$$g(S_{M,j}) = \sum_{p=1}^{w/s} (p^{th}\ segment\ of\ S_{M,j}) + \sum_{q=1}^{w/s-1} q^{th}\ joint\ of\ S_{M,j}$$

(Eq. 7.3)

Dot product of Eq.7.2 and Eq.7.3 is then

$$g(S_{H,i}) \cdot g(S_{M,j})$$
$$= \left(\sum_{p_1=1}^{w/s} (p_1^{th}\ segment\ of\ S_{H,i}) + \sum_{q_1=1}^{w/s-1} q_1^{th}\ joint\ of\ S_{H,i} \right)$$
$$\cdot \left(\sum_{p_2=1}^{w/s} (p_2^{th}\ segment\ of\ S_{M,j}) + \sum_{q_2=1}^{w/s-1} q_2^{th}\ joint\ of\ S_{M,j} \right)$$

$$= \sum_{p_1=1}^{w/s} \sum_{p_2=1}^{w/s} \left(p_1^{th}\ segment\ of\ S_{H,i}\right) \cdot \left(p_2^{th}\ segment\ of\ S_{M,j}\right)$$

$$+ \sum_{p_1=1}^{w/s} \sum_{q_2=1}^{w/s-1} \left(p_1^{th}\ segment\ of\ S_{H,i}\right) \cdot \left(q_2^{th}\ joint\ of\ S_{M,j}\right)$$

$$+ \sum_{q_1=1}^{w/s-1} \sum_{p_2=1}^{w/s} \left(q_1^{th}\ joint\ of\ S_{H,i}\right) \cdot \left(p_2^{th}\ segment\ of\ S_{M,j}\right)$$

$$+ \sum_{q_1=1}^{w/s-1} \sum_{q_2=1}^{w/s-1} \left(q_1^{th}\ joint\ of\ S_{H,i}\right) \cdot \left(q_2^{th}\ joint\ of\ S_{M,j}\right)$$

(Eq. 7.4)

, where each of the dot products of gapped-kmer vector is computed using Eq.4.5. The dot products in Eq 7.4 can then be partitioned into four subsets (according to how they grouped into the four double series in Eq.7.4) and be organized into a 3D hypermatrix.

$$P^{ij} \in \mathbb{N}^{(w/s)\times(w/s)\times 4}$$

$$P^{ij}[q_1, q_2, 1] = g\left(q_1^{th}\ segment\ of\ S_{H,i}\right) \cdot g\left(q_2^{th}\ segment\ of\ S_{M,j}\right)$$

$$P^{ij}[q_1, q_2, 2] = g\left(q_1^{th}\ segment\ of\ S_{H,i}\right) \cdot g\left(q_2^{th}\ joint\ of\ S_{M,j}\right)$$

$$P^{ij}[q_1, q_2, 3] = g\left(q_1^{th}\ joint\ of\ S_{H,i}\right) \cdot g\left(q_2^{th}\ segment\ of\ S_{M,j}\right)$$

$$P^{ij}[q_1, q_2, 4] = g\left(q_1^{th}\ joint\ of\ S_{H,i}\right) \cdot g\left(q_2^{th}\ joint\ of\ S_{M,j}\right)$$

Then, Eq.7.4 equivalent to the equation below (Eq.7.5).

$$g\left(S_{H,i}\right) \cdot g\left(S_{M,j}\right) = \sum_{q_1=1}^{w/s} \sum_{q_2=1}^{w/s} \sum_{k=1}^{4} P^{ij}[q_1, q_2, k] \cdot \mathbb{1}(q_1, q_2, k)$$

(Eq. 7.5)

Where $\mathbb{1}(\mathrm{q}_1, \mathrm{q}_2, k)$ is a Boolean function enforcing the index ranges of the double series in Eq.7.4.

$$\mathbb{1}(\mathrm{q}_1, \mathrm{q}_2, k)\begin{cases} 1 & & ,k=1 \\ 1 & iff\ q_2 < w/s & ,k=2 \\ 1 & iff\ q_1 < w/s & ,k=3 \\ 1 & iff\ q_1, q_2 < w/s & ,k=4 \end{cases}$$

Note that the entries of P^{ij} are redundantly shared with P^{xy} from neighboring pairs of sliding windows (e.g., $P^{(i-1)j}$,$P^{i(j-1)}$,$P^{(i-1)(j-1)}$, etc.). Define matrix P, which has every P^{ij} as its submatrices and non-redundantly contains all the entries of P^{ij} for all i^{th} and j^{th} sliding windows.

$$P \in N^{(|S_H|/s)\times(|S_M|/s)\times 4}$$

$$P[(i-1)+q_1, (j-1)+q_2, :] = P^{i,j}[q_1, q_2, :]\ ,\ \text{where } 1 \le q_1, q_2 \le w/s$$

which leads to:

$$\mathrm{P}[(\mathrm{i}-1)+\mathrm{q}_1, (\mathrm{j}-1)+\mathrm{q}_2, 1] = g\left(q_1^{th}\ segment\ of\ S_{H,i}\right) \cdot g(q_2^{th} segment\ of\ S_{M,j})$$

$$\mathrm{P}[(\mathrm{i}-1)+\mathrm{q}_1, (\mathrm{j}-1)+\mathrm{q}_2, 2] = g\left(q_1^{th}\ segment\ of\ S_{H,i}\right) \cdot g(q_2^{th} joint\ of\ S_{M,j})$$

$$\mathrm{P}[(\mathrm{i}-1)+\mathrm{q}_1, (\mathrm{j}-1)+\mathrm{q}_2, 3] = g\left(q_1^{th}\ joint\ of\ S_{H,i}\right) \cdot g(q_2^{th} segment\ of\ S_{M,j})$$

$$\mathrm{P}[(\mathrm{i}-1)+\mathrm{q}_1, (\mathrm{j}-1)+\mathrm{q}_2, 4] = g\left(q_1^{th}\ joint\ of\ S_{H,i}\right) \cdot g(q_2^{th} joint\ of\ S_{M,j})$$

Finally, Eq.7.5 is equivalently and more compactly expressed as:

$$g\left(S_{H,i}\right) \cdot\ g\left(S_{M,j}\right) = \sum_{q_1=1}^{w/s} \sum_{q_2=1}^{w/s} \sum_{k=1}^{4} P[(i-1)+q_1, (j-1)+q_2, k] \cdot \mathbb{1}(q_1, q_2, k)$$

(Eq. 7.5)

The matrix P contains all necessary and sufficient information for computing $G_{S_H,S_M}[i,j]$ for all pairs of i^{th} and j^{th} sliding windows of S_H and S_M. It contains gapped-kmer similarities

of all pairs of atomic fragments of S_H and S_M that can be combinatorially summed to compute each element of G_{S_H,S_M}. Each element in P is non-redundantly computed and stored. P is initialized to be a matrix of -1, and its elements are replaced with a non-negative number while computing G. For example, $P[1,2,1]$ is needed both for computing $g(S_{H,1}, S_{M,1})$ (via $i, j, q_1, q_2 = 1,1,1,2$) and for computing $g(S_{H,1}, S_{M,2})$ (via $i, j, q\_1, q\_2 = 1,2,1,1$), but it is computed once while computing $g(S_{H,1}) \cdot g(S_{M,1})$. All elements of P are computed once and stored during the course of computing G, and they are accessed when they are needed for computing other elements of G. This optimization makes the time-complexity of gkm-align nearly independent of the sliding window width (w).

Hence, Eq 7.5 gives gkm-align algorithm the time-complexity independent of both l (gapped-kmer width; through SIMD) and w (window width). It is dependent on the sizes of the genomes and sizes of the human-mouse syntenic blocks (**Figure 7.1.D**; **chapter 7.4**).

7.3 Finding the optimal alignment path of maximum gapped-kmer similarity using matrix G

After computing matrix G, as described in **chapter 7.2**, we can now compute the optimal collinear alignment path spanning the human and mouse genomic loci at the resolution of sliding windows. Resulting alignment path is a collinear sequence of pairs of human and mouse genomic windows that most likely contain conserved DNA elements inherited from the human-mouse common ancestor.

Provided with matrix $G \in [0,1]^{m \times n}$ that contains pairwise sequence-similarities of every pair of sliding windows in S_H and S_M, we define a collinear path of sliding windows as a sequence of matrix indices. The matrix indices are eventually converted to human-

mouse genomic coordinates for gkm-align output. $P = \left((i_1, j_1), (i_2, j_2) \ldots, (i_{|P|}, j_{|P|})\right)$, such that

i. $(i_1, j_1) = (1,1)$

ii. $(i_{|P|}, j_{|P|}) = (m, n)$

iii. $(i_{k+1}, j_{k+1}) - (i_k, j_k) \in \{(1,1), (1,0), (0,1)\}$

iv. $p \geq q \rightarrow i_p \geq i_q \; and \; j_p \geq j_q$

where $(i_{k+1}, j_{k+1}) - (i_k, j_k) \in \{(1,0), (0,1)\}$ indicates insertion or deletion events in S_H or S_M relative to each other.

To find the most likely path that describes the evolutionary history between S_H and S_M, we use a variant of the Smith-Waterman algorithm. To find the optimal path P, we score all possible paths satisfying [i-iv] as follows:

$$f(P) = \sum_{q=1}^{|P|} G\big(P(i)\big) - \alpha \cdot I$$

where I is the number of indel events in P.

$$I = \left|\left\{k \in 1 \ldots |P| - 1 : (i_{k+1}, j_{k+1}) - (i_k, j_k) \in \{(1,0), (0,1)\}\right\}\right|$$

It is difficult to choose a stable value for α since it depends on the parameters used for computing the sequence similarities, such as the width of the sliding windows (w) and l and k for the gapped-kmer definition. To choose α independent of (w, l, k), we Z-transform matrix G so that

$$f(P) = \sum_{q=1}^{|P|} \frac{G\big(P(i)\big) - \mu_G}{\sigma_G} - \alpha \cdot I$$

(Eq.7.6)

Z-transforming also has an additional advantage of normalizing sequence similarities relative to local background sequence similarities, which are variable by loci, for example, by GC contents. Globally optimal P for a given G is then efficiently identified using dynamic programming.

$$optimal\ alignment\ path\ spanning\ S_H\ and\ S_M$$

$$= argmax_P\ f(P)$$

$$= argmax_P \sum_{q=1}^{|P|} \frac{G\big(P(i)\big) - \mu_G}{\sigma_G} - \alpha \cdot I$$

$$= argmax_P \sum_{q=1}^{|P|} \frac{G\big(P(i)\big) - \mu_G}{\sigma_G}\ ,if\ \alpha = 0$$

(Eq.7.7)

7.4 Whole-genome extension of gkm-align

Chapter 7.1-7.4 explained how local gkm-alignment is performed Figure 7.1A-C given that the human-mouse syntenic blocks of interest are specified (visualized as the rectangles in **Figure 7.1D**). Here, we discuss how gkm-align is extended genome-wide. The essence of whole-genome extension is the identification of human-mouse syntenic blocks. Most conventional genome-alignment methods[21,22,190] identify syntenic blocks by (1) identifying short sequence matches between human and mouse genomes (2) and chaining the matches based on their genomic coordinates. Gkm-align adopts this idea.

First, we obtained short sequence matches (\textit{seed hits}) between human and mouse using LASTZ (v1.04.04; options: --notransition --step=1 --gfextend --nochain --

nogapped) (visualized by the dots in **Figure 7.1D**). Then, to focus on conserved gene regulatory circuits, we filtered out seed hits that are not located within any human/mouse syntenic intergenic locus identified using the ENCODE gene ortholog list[100].

7.5 Enhanced discovery of conserved enhancers by incorporating cell-specific regulatory vocabularies

We can further enhance discovery of orthologous enhancer pairs using sequence-based machine learning models. Gkm-SVM models can distinguish enhancers from random genomic background, and their prediction values for each human and mouse elements can be weighted into sequence similarity matrix G to enhance alignment.

Recall that sequence similarity between i^{th} and j^{th} slidindg windows of sequences S_H and S_M (denoted as $S_{H,i}$, $S_{M,j}$) is computed as

$$G_{S_H,S_M}[i,j] = \frac{g(S_{H,i}) \cdot g(S_{M,j})}{\|g(S_{H,i})\| \cdot \|g(S_{M,j})\|}$$

Then, gkm-SVM weighted sequence similarity is computed as

$$G_{S_H,S_M}[i,j] \leftarrow G_{S_H,S_M}[i,j]^{(1-c)} \cdot \left(P_M(S_{H,i}) \cdot P_H(S_{M,j})\right)^c$$

(Eq.7.8)

where P_M and P_H are enhancer prediction functions ($P: sequence \rightarrow [0,1]$) trained on mouse (M) and human (H) enhancers. Higher $c \in [0,1]$ indicates higher influence of enhancer predictions on alignment.

One limitation of directly incorporating gkm-SVM predictions to gkm-align is that these values are not bounded, limiting interpretable application for genome alignment. Instead, we transform the prediction values to range between 0 and 1 by computing the posterior probability that a sequence is an enhancer. This is done by computing gkm-SVM score distributions for enhancers and non-enhancers (i.e., positive and negative training sets used for gkm-SVM enhancer training; prediction scores computed with 5-fold cross validation)

$$\mathrm{mu}_{(+)} = \mathrm{E}\left(\mathrm{pred}\left(\mathrm{X}_{\mathrm{pos}}\right)\right), \sigma^2_{(+)} = \mathrm{Var}\left(\mathrm{pred}\left(\mathrm{X}_{\mathrm{pos}}\right)\right)$$

$$\mathrm{mu}_{(-)} = \mathrm{E}\left(\mathrm{pred}\left(\mathrm{X}_{\mathrm{neg}}\right)\right), \sigma^2_{(-)} = \mathrm{Var}\left(\mathrm{pred}\left(\mathrm{X}_{\mathrm{neg}}\right)\right)$$

Assuming that priors for finding positive and negative sequences are equal, we compute posterior gkm-SVM prediction score of a given sequence X with

$$\mathrm{P}_{\mathrm{gkm}}(\mathrm{X}) = \mathrm{P}\left(\mathrm{X} \in \mathrm{X}_{\mathrm{pos}} \middle| \mathrm{pred}(\mathrm{X})\right) = \frac{\mathrm{P}_{(+)}}{\mathrm{P}_{(+)} + \mathrm{P}_{(-)}} \in [0,1],$$

$$\mathrm{P}_{(+)} = \frac{1}{\sqrt{2\pi\sigma^2_{(+)}}} \exp\left(\frac{\mathrm{pred}(\mathrm{X}) - \mu_{(+)}}{2\sigma^2_{(+)}}\right)$$

$$\mathrm{P}_{(-)} = \frac{1}{\sqrt{2\pi\sigma^2_{(-)}}} \exp\left(\frac{\mathrm{pred}(\mathrm{X}) - \mu_{(-)}}{2\sigma^2_{(-)}}\right)$$

$pred(X)$ is computed using Eq.4.7, summing up kmer-weights of every kmer contained in X.

7.6 Detecting and masking repetitive elements using gkm-SVM

Without masking repetitive DNA elements, Matrix G is riddled with sequence similarities coming from high-entropy repetitive elements that are prevalent across the human and mouse genomes (mostly simple low-complexity repeats; **Fig 6.2**). Hence, a method to underweight sequence similarities coming from repetitive elements is crucial for reliable genome alignment.

One intuitive solution is to use principal component analysis on the rows and columns of G to automatically detect elements in S_H that are similar to many elements in S_M and vice versa. However, such a method is computationally too inefficient to apply genome-wide. Instead, we can again use sequence-based machine learning methods (e.g. gkm-SVM) to learn patterns of DNA sequences that are highly populated in the human and mouse genomes, and use the model to detect and filter out repetitive elements in linear time.

We first train gkm-SVM, separately for human and mouse, on randomly sampled genomic windows (300 base pairs wide) outside regulatory regions (e.g., union of DHS across all ENCODE experiments) against randomly generated sequences such that each base pair has equal probabilities of being A,C,G, or T. This results in a kmer-weight vector, where kmers with high weight are significantly enriched in the genome more than expected by chance (**Fig 6.2**; e.g., ATATATATATA).

In order to detect base pairs that highly contribute to repetitiveness observed in matrix G, we compute repeat scores of each base pair in the genome by summing up kmer weights of kmers that overlap with the base pair. For i^{th} base pair in sequence S, its repeat score is computed as

$$r(S[i]) = \sum_{j=i-k+1}^{i} w_{S(j,k)},$$

where k = kmer size. Using this function, we can compute repeat scores for each base pair across the entire genome, and mask base pairs that receive repeat scores above a certain threshold, determined by the percentage of the genomic base pairs that we target to mask. For gkm-align, we used a repeat score threshold that would mask approximately 10 percent of the human and mouse genomes. Base pairs that surpass the estimated threshold are masked by replacement with random characters. They are replaced with ASCII characters excluding (A, C, G, T) in order to nearly eliminate the chance that the newly introduced characters contribute any artificial sequence similarities. After repeat masking, we compute matrix G as described in chapter 7.2, compute optimal alignment paths using G (chapter 7.3), and repeat this process across every human-mouse syntenic blocks (chapter 7.4) for whole-genome alignment.

7.7 Discussion

In this chapter, I provided detailed description of the gkm-align algorithm, including optimization techniques I applied to reduce its time-complexity. The source code (C++) of gkm-align is accessible through the github page: https://github.com/oh-jinwoo94/gkm-align. I also included a detailed instruction of how to use gkm-align for genome-alignment, which can be used to map enhancers conserved between human and mouse. Gkm-align may be used to map DNA elements other than enhancers, although its performance has been evaluated only in the context of regulatory elements.

Chapter 8

Discussion

The initial drafts of the human genome revealed that only around 1% of the genome codes protein-coding genes[130,131], with largely unknown biological function for the rest of the genome. Elucidating the role of the non-coding genome has been a major focus of the ENCODE (Encyclopedia of DNA Elements) consortium[32,82]. Over the past two decades, the consortium has mapped around 1 million human and mouse putative cis-regulatory elements using biomarkers associated with regulatory activities, which has revealed that a large fraction of disease-associated variants are enriched in regulatory elements[1]. A growing body of work is revealing that variation in intergenic regulatory elements contributes to cancer progression, heterogeneity, severity, and response to treatment[191–194]. Despite the significant medical implication, mapping individual enhancers to their biological functions has been difficult. For example, identifying an enhancer's target promoter is difficult as enhancers are often distal to their target promoters, often skipping multiple genes[12–14]. Also, many human genes are regulated by complex networks of transcription factors (TF) and by multiple enhancers, and it is unclear how such network properties affect phenotypes and how their disruptions cause disease[8,15,16]. Further, testing human regulatory elements through mouse models is challenging, especially for distal enhancers. Due to both rapid evolution and sequence complexities of enhancers, only a small fraction of human enhancers is mappable to the mouse genome with conserved regulatory activities using conventional computational methods[20–22,159]. Overcoming the challenges imposed by the regulatory and sequence complexities of enhancers has been our main scientific pursuit.

First, we had extensive collaboration through the ENCODE consortium to functionally characterize diverse gene regulatory elements using CRISPR, and have generated the largest public database for CRISPR screens to date[9] (**chapter 3**). We epigenetically perturbed a total of 20 megabases of the human genome across 16 human cell lines and quantified the cellular phenotypic impacts. The large-scale genomic perturbations were achieved by utilizing pooled CRISPRi screen methods, which deliver to cell populations numerous guide RNAs targeting many putative regulatory elements in parallel. The delivered guide RNAs recruit catalytically-dead Cas9 proteins attached with a repressor domain (dCas9-KRAB; CRISPRi[17]) to the target loci. The infected cells are sorted based on their perturbed phenotypes such as gene expression[103,104] and proliferation rate[114,132], and subsequently sequenced for guide enrichment. Through quantitative analysis of the CRISPR database, we showed that CRISPR screens reproducibly generate phenotype-perturbation signals and accurately recover well-characterized distal enhancers (e.g., GATA1 enhancers in K562 cancer cells). Using a known set of regulatory elements, we quantitatively modeled CRISPR experiments to generate an experimental guideline for conducting high-quality CRISPR screens with minimal sequencing depth and cell coverage. Lastly, we discovered previously unreported properties of CRISPR perturbation, including the strand-specific effect of perturbing gene-bodies with CRISPRi. These findings have significant implications for designing future CRISPR screens of regulatory elements. Together, we presented a public database of CRISPR screens and also uncovered important properties of CRISPR screens that will facilitate future production of CRISPR screens for functional characterization of gene regulatory elements.

[9] Oh, J.W., Yao, Tycko, Bounds, Gosai, Lataniotis, . *et al.* Multicenter integrated analysis of noncoding CRISPRi screens. Nature Methods (2024)

Next, we conducted a CRISPRi screen to identify distal enhancers responsible for driving the differentiation of embryonic stem cells (ESC) into definitive endoderm (DE)[10] (**chapter 3**). We screened 394 putative enhancers near the core TFs of ESC (OCT4, SOX2, NANOG) and DE (EOMES, SOX17, GATA6, MIXL1), and quantified the degree to which perturbation of each putative enhancer reduces the ESC-DE differentiation rate. We found that detection of enhancers involved in cell differentiation is temporally sensitive. At mid-transition timepoint (36 hours; DE-36h), we identified 29 enhancers that significantly reduced the cell-transition rate when perturbed with CRISPRi. However, with sufficient amount of time of DE state induction (Activin-A treatment for 72 hours; DE-72h), most of the CRISPRi perturbations targeting the 29 enhancers had weak effect on ESC-DE differentiation rate. This observation implies that CRISPRi perturbation of the enhancers leads to only a time-delay in cell-state transition, with little impact post-transition. Based on this observation, we developed mathematical models that closely mimic and explain the cell-transition dynamics revealed through the CRISPR screen. Further, using the enhancer CRISPR screen data, we developed a CTCF-loop based machine learning method for predicting enhancer-promoter interaction, which outperformed previous methods based on Hi-C contact frequency[104]. Together, we discovered the regulatory process driving the ESC-DE differentiation, and also presented a generalizable strategy that combines CRISPR functional characterization and mathematical modeling for discovering enhancers and elucidating mechanisms of cell-state transition.

Lastly, to facilitate functional characterization of human enhancers through mouse models, we developed a novel genome-alignment algorithm for improved mapping of

[10] Luo, R., Yan, J., Oh, J.W. *et al.* Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions. Nature Genetics 55, 1336–1346 (2023).

conserved distal enhancers[11] (**chapter 4-7**). Our previously developed machine learning model, which uses gapped-kmer sequence features (gkm-SVM[3,23]), have shown to effectively model enhancer sequences[30] and accurately predict gene expression[28]. Gkm-SVM encodes sequence motifs prevalent in enhancers in a kmer-weight vector, where kmers associated with TF binding are assigned higher weights. Gkm-SVM kmer-weight vectors, or *enhancer vocabularies*, are highly conserved between human and mouse in a highly cell-specific manner, implying that cell-specific TF expression and binding patterns associated with enhancers are highly (**chapter 4**). In contrast, enhancers appear to be weakly conserved, as orthologous mouse enhancers mapped from human enhancers using conventional genome-alignment algorithms often lack conserved regulatory activities[20–22,159] (**chapter 5**). To improve the identification of conserved enhancers, we developed a novel genome-alignment algorithm, *gkm-align* using gapped-kmer sequence features (**chapter 7**). To evaluate gkm-align, we generated a list of 45 orthologous human and mouse cell/tissues from >1,000 ENCODE DNase-seq experiments, and showed that *gkm-align* can discover more than 20,000 conserved enhancers previously unidentifiable with conventional computational methods (**chapter 6**). For example, we identified a conserved intronic enhancer of OTUD7A, a gene that is associated with schizophrenia and epilepsy in human[181,195] and causes cortical neuron deformation and seizure-events in OTUD7A knockout mice[182]. During the evaluation process, we observed a surprisingly high level of cell-specificity in enhancer conservation, largely explainable by cell-specific association with transposable elements. The cell-specific pattern of enhancer conservation is consistent with the cell-specific pattern of conservation in gene expression between human and mouse. Despite the addition of novel conserved enhancers using gkm-align, the overall enhancer conservation levels remained low. This observation was

[11] Oh, J.W., and Beer, MA. Gapped-kmer sequence modeling robustly identifies regulatory vocabularies and distal enhancers conserved between evolutionarily distant mammals. *bioRxiv* (2023; under review)

also confirmed using alignment-free analysis, suggesting that rapid enhancer evolution is a fundamental property of mammalian evolution. Using *gkm-align*, we published an expanded catalogue of conserved enhancers, which we believe will streamline functional characterization of human enhancers, and I aspire to contribute to advancements in the diagnosis and treatment of regulatory diseases through our research efforts.

Bibliography

1. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
3. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
4. Beer, M. A. Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* (2017).
5. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
6. Ludwig, M. Z. Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.* **12**, 634–639 (2002).
7. Rebeiz, M. & Tsiantis, M. Enhancer evolution and the origins of morphological novelty. *Curr. Opin. Genet. Dev.* **45**, 115–123 (2017).
8. Luo, R. *et al.* Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions. *Nat. Genet.* (2023) doi:10.1038/s41588-023-01450-7.
9. Yao, D. *et al.* Multicenter integrated analysis of noncoding CRISPRi screens. *Nat. Methods* (2024) doi:10.1038/s41592-024-02216-7.
10. Oh, J. W. & Beer, M. A. Gapped-kmer sequence modeling robustly identifies regulatory vocabularies and distal enhancers conserved between evolutionarily distant mammals. *bioRxiv* 2023.10.06.561128 (2023) doi:10.1101/2023.10.06.561128.

11. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).

12. Anderson, E., Devenney, P. S., Hill, R. E. & Lettice, L. A. Mapping the Shh long-range regulatory domain. *Development* **141**, 3934–3943 (2014).

13. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).

14. Martin, P. *et al.* Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* **6**, 10069 (2015).

15. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).

16. Kvon, E. Z., Waymack, R., Gad, M. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nat. Rev. Genet.* **22**, 324–336 (2021).

17. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

18. McClymont, S. A. *et al.* Parkinson-Associated SNCA Enhancer Variants Revealed by Open Chromatin in Mouse Dopamine Neurons. *Am. J. Hum. Genet.* **103**, 874–892 (2018).

19. Symmons, O. *et al.* The shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Dev. Cell* **39**, 529–543 (2016).

20. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).

21. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 11484–11489 (2003).

22. Harris, R. S. Improved pairwise alignment of genomic DNA. (2007).

23. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).

24. Beer, M. A., Shigaki, D. & Huangfu, D. Enhancer Predictions and Genome-Wide Regulatory Circuits. *Annu. Rev. Genomics Hum. Genet.* **21**, 37–54 (2020).

25. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).

26. de Souza, F. S. J. & Franchini, L. F. Exaptation of Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always Strong? *Mol. Biol.* (2013).

27. Senft, A. D. & Macfarlan, T. S. Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021).

28. Shigaki, D. *et al.* Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.* **40**, 1280–1291 (2019).

29. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).

30. Yan, J. *et al.* Systematic analysis of binding of transcription factors to noncoding variants. *Nature* **591**, 147–151 (2021).

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

32. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

33. Giorgetti, L. *et al.* Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. *Mol. Cell* **37**, 418–428 (2010).

34. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).

35. Soutourina, J. Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol.* **19**, 262–274 (2018).

36. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17**, 487–500 (2016).

37. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214 (1996).

38. Senger, K. *et al.* Immunity regulatory DNAs share common organizational features in Drosophila. *Mol. Cell* **13**, 19–32 (2004).

39. Swanson, C. I., Evans, N. C. & Barolo, S. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev. Cell* **18**, 359–370 (2010).

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

41. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet.* **4**, e1000106 (2008).

42. Moses, A. M. *et al.* Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.* **2**, e130 (2006).

43. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890–898 (2005).

44. Lin, Y. S., Carey, M., Ptashne, M. & Green, M. R. How different eukaryotic transcriptional activators can cooperate promiscuously. *Nature* **345**, 359–361 (1990).

45. Liberman, L. M. & Stathopoulos, A. Design flexibility in cis-regulatory control of gene expression: Synthetic and comparative evidence. *Dev. Biol.* **327**, 578–589 (2009).

46. Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65–70 (2009).

47. Kreimer, A. *et al.* Predicting gene expression in massively parallel reporter assays: A comparative study. *Hum. Mutat.* **38**, 1240–1250 (2017).

48. Patel, Z. M. & Hughes, T. R. Global properties of regulatory sequences are predicted by transcription factor recognition mechanisms. *Genome Biol.* **22**, 285 (2021).

49. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).

50. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

51. Stormo, G. D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.* **11**, 751–760 (2010).

52. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016).

53. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

54. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

55. Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 22534–22539 (2010).

56. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions. *Nature* **604**, 571–577 (2022).

57. Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).

58. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

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

60. Udenberg, G. Formation of chromosomal domains by loop extrusion. *Cell Rep* **15**, 2038–2049 (2016).

61. Liu, X. S. *et al.* Editing DNA methylation in the mammalian genome. *Cell* **167**, 233-247.e17 (2016).

62. Ji, X. *et al.* 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18**, 262–275 (2016).

63. Xi, W. & Beer, M. A. Loop competition and extrusion model predicts CTCF interaction specificity. *Nat. Commun.* **12**, 1046 (2021).

64. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

65. Giorgetti, L. *et al.* Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–963 (2014).

66. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).

67. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).

68. Tuan, D., Solomon, W., Li, Q. & London, I. M. The “beta-like-globin” gene domain in human erythroid cells. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 6384–6388 (1985).

69. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).

70. Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J.* **7**, 1395–1402 (1988).

71. Lee, D. Y., Hayes, J. J., Pruss, D. & Wolffe, A. P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73–84 (1993).

72. Hebbes, T. R., Clayton, A. L., Thorne, A. W. & Crane-Robinson, C. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. *EMBO J.* **13**, 1823–1830 (1994).

73. Serfling, E., Jasin, M. & Schaffner, W. Enhancers and eukaryotic gene transcription. *Trends Genet.* **1**, 224–230 (1985).

74. Ikuta, T. & Kan, Y. W. In vivo protein-DNA interactions at the beta-globin gene locus. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 10188–10192 (1991).

75. Forsberg, M. & Westin, G. Enhancer activation by a single type of transcription factor shows cell type dependence. *EMBO J.* **10**, 2543–2551 (1991).

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

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

78. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

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

80. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).

81. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

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

83. Maston, G. A., Landt, S. G., Snyder, M. & Green, M. R. Characterization of enhancer function from genome-wide analyses. *Annu. Rev. Genomics Hum. Genet.* **13**, 29–57 (2012).

84. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).

85. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).

86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

87. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

88. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

89. Hatzis, P. & Talianidis, I. Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol. Cell* **10**, 1467–1477 (2002).

90. Wang, Q., Carroll, J. S. & Brown, M. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol. Cell* **19**, 631–642 (2005).

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

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

93. Roller, M. *et al.* LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol.* **22**, 62 (2021).

94. Wang, Z. & Ren, B. Role of H3K4 monomethylation in gene regulation. *Curr. Opin. Genet. Dev.* **84**, 102153 (2024).

95. Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science* **328**, 1161–1164 (2010).

96. Sung, M.-H., Guertin, M. J., Baek, S. & Hager, G. L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell* **56**, 275–285 (2014).

97. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, db.prot5384 (2010).

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

99. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

100. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).

101. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR-Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).
102. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
103. Reilly, S. K. *et al.* Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR-FlowFISH. *Nat. Genet.* **53**, 1166–1176 (2021).
104. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
105. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
106. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat. Genet.* **40**, 322–328 (2008).
107. Sharpe, J. *et al.* Identification of sonic hedgehog as a candidate gene responsible for the polydactylous mouse mutant Sasquatch. *Curr. Biol.* **9**, 97–100 (1999).
108. Danna, K. & Nathans, D. Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. 1971. *Rev. Med. Virol.* **9**, 75–81 (1999).
109. Liu, Q., Segal, D. J., Ghiara, J. B. & Barbas, C. F., 3rd. Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 5525–5530 (1997).
110. Christian, M. *et al.* Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757–761 (2010).
111. DI Felice, F., Micheli, G. & Camilloni, G. Restriction enzymes and their use in molecular biology: An overview. *J. Biosci.* **44**, (2019).
112. Hart, T. *et al.* High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).

113. Li, Q. V. *et al.* Genome-scale screens identify JNK-JUN signaling as a barrier for pluripotency exit and endoderm differentiation. *Nat. Genet.* **51**, 999–1010 (2019).

114. Gilbert, L. A. *et al.* Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).

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

116. Jansen, R., van Embden, J. D. A., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).

117. van Embden, J. D. *et al.* Genetic variation and evolutionary origin of the direct repeat locus of Mycobacterium tuberculosis complex bacteria. *J. Bacteriol.* **182**, 2393–2401 (2000).

118. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).

119. Hille, F. & Charpentier, E. CRISPR-Cas: biology, mechanisms and relevance. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150496 (2016).

120. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

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

122. Dixit, A. *et al.* Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853-1866.e17 (2016).

123. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

124. Margolin, J. F. *et al.* Krüppel-associated boxes are potent transcriptional repression domains. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 4509–4513 (1994).

125. Kearns, N. A. *et al.* Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. *Development* **141**, 219–223 (2014).

126. Leonetti, M. D., Sekine, S., Kamiyama, D., Weissman, J. S. & Huang, B. A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E3501-8 (2016).

127. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).

128. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

129. Perez, A. R. *et al.* GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35**, 347–349 (2017).

130. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

131. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

132. Tycko, J. *et al.* Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nat. Commun.* **10**, 4063 (2019).

133. Shi, Z.-D. *et al.* Genome editing in hPSCs reveals GATA6 haploinsufficiency and a genetic interaction with GATA4 in human pancreatic development. *Cell Stem Cell* **20**, 675-688.e6 (2017).

134. Gillespie, D. T. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55 (2007).

135. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).

136. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

137. Chang, C.-C. & Lin, C.-J. LIBSVM. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).

138. Edsall, L. E. *et al.* Evaluating chromatin accessibility differences across multiple primate species using a joint modeling approach. *Genome Biol. Evol.* **11**, 3035–3053 (2019).

139. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).

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

141. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014).

142. ENCODE 4 flagship [ENC4103].

143. Gowan, K. *et al.* Crossinhibitory activities of Ngn1 and Math1 allow specification of distinct dorsal interneurons. *Neuron* **31**, 219–232 (2001).

144. Flora, A., Garcia, J. J., Thaller, C. & Zoghbi, H. Y. The E-protein Tcf4 interacts with Math1 to regulate differentiation of a specific subset of neuronal progenitors. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 15382–15387 (2007).

145. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).

146. Harris, H. K. *et al.* Disruption of RFX family transcription factors causes autism, attention-deficit/hyperactivity disorder, intellectual disability, and dysregulated behavior. *Genet. Med.* **23**, 1028–1040 (2021).

147. Le Coz, C. *et al.* Constrained chromatin accessibility in PU.1-mutated agammaglobulinemia patients. *J. Exp. Med.* **218**, (2021).

148. Chandra, V. *et al.* Multidomain integration in the structure of the HNF-4α nuclear receptor complex. *Nature* **495**, 394–398 (2013).

149. Lin, S. *et al.* Comparison of the transcriptional landscapes between human and mouse tissues. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 17224–17229 (2014).

150. Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat. Commun.* **6**, 5903 (2015).

151. Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434–438 (2008).

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

153. Attanasio, C. *et al.* Fine tuning of craniofacial morphology by distant-acting enhancers. *Science* **342**, 1241006 (2013).

154. Shin, J. Y. *et al.* Epigenetic activation and memory at a TGFB2 enhancer in systemic sclerosis. *Sci. Transl. Med.* **11**, eaaw0790 (2019).

155. Hardison, R. C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.* **13**, 469–483 (2012).

156. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

157. Sullivan, P. F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).

158. Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371–375 (2014).

159. Kuhn, R. M. *et al.* The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.* **37**, D755-61 (2009).

160. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4.10 (2004).

161. Xi, W. & Beer, M. A. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput. Biol.* **14**, e1006625 (2018).

162. Teufel, A. *et al.* Comparison of Gene Expression Patterns Between Mouse Models of Nonalcoholic Fatty Liver Disease and Liver Tissues From Patients. *Gastroenterology* **151**, 513-525.e0 (2016).

163. Jiang, C. *et al.* Comparative Transcriptomics Analyses in Livers of Mice, Humans, and Humanized Mice Define Human-Specific Gene Networks. *Cells* **9**, (2020).

164. Du, A. Y., Chobirko, J. D., Zhuo, X., Feschotte, C. & Wang, T. Regulatory Transposable Elements in the Encyclopedia of DNA Elements. *bioRxiv* 2023.09.05.556380 (2023) doi:10.1101/2023.09.05.556380.

165. Kazazian, H. H., Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).

166. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).

167. Pontis, J. *et al.* Primate-specific transposable elements shape transcriptional networks during human development. *Nat. Commun.* **13**, 7178 (2022).

168. Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* **23**, 481–497 (2022).

169. Breschi, A. *et al.* Gene-specific patterns of expression variation across organs and species. *Genome Biol.* **17**, 151 (2016).

170. IGVF. The Impact of Genomic Variation on Function (IGVF) Consortium. *ArXiv* (2023) doi:10.1101/2023.03.28.533945.

171. Ghandi, M., Mohammad-Noori, M. & Beer, M. A. Robust k-mer frequency estimation using gapped k-mers. *J. Math. Biol.* **69**, 469–500 (2014).

172. Amanchy, R. *et al.* Identification of novel phosphorylation motifs through an integrative computational and experimental analysis of the human phosphoproteome. *J. Proteomics Bioinform.* **4**, 22–35 (2011).

173. van Helden, J. Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics* **20**, 399–406 (2004).

174. Kantorovitz, M. R., Robinson, G. E. & Sinha, S. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* **23**, i249-55 (2007).

175. Göke, J., Schulz, M. H., Lasserre, J. & Vingron, M. Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics* **28**, 656–663 (2012).

176. Zhang, Z., Raghavachari, B., Hardison, R. C. & Miller, W. Chaining multiple-alignment blocks. *J. Comput. Biol.* **1**, 217–226 (1994).

177. Philipsen, S. & Hardison, R. C. Evolution of hemoglobin loci and their regulatory elements. *Blood Cells Mol. Dis.* **70**, 2–12 (2018).

178. Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**, 975–985 (1987).

179. Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res.* **19**, 2172–2184 (2009).

180. Lowther, C. *et al.* Delineating the 15q13.3 microdeletion phenotype: a case series and comprehensive review of the literature. *Genet. Med.* **17**, 149–157 (2015).

181. Uddin, M. *et al.* OTUD7A Regulates Neurodevelopmental Phenotypes in the 15q13.3 Microdeletion Syndrome. *Am. J. Hum. Genet.* **102**, 278–295 (2018).

182. Yin, J. *et al.* Otud7a Knockout Mice Recapitulate Many Neurological Features of 15q13.3 Microdeletion Syndrome. *Am. J. Hum. Genet.* **102**, 296–308 (2018).

183. Negrisolo, S. *et al.* PAX2 gene mutations in pediatric and young adult transplant recipients: kidney and urinary tract malformations without ocular anomalies. *Clin. Genet.* **80**, 581–585 (2011).

184. Shukla, A., Narayanan, D. L., Asher, U. & Girisha, K. M. A novel bi-allelic loss-of-function variant in MYOD1: Further evidence for gene-disease association and phenotypic variability in MYOD1-related myopathy. *Clin. Genet.* **96**, 276–277 (2019).

185. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).

186. Lazarian, G. *et al.* A hotspot mutation in transcription factor IKZF3 drives B cell neoplasia via transcriptional dysregulation. *Cancer Cell* **39**, 380-393.e8 (2021).

187. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).

188. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

189. Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01793-w.

190. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999).

191. Xing, M. *et al.* Genomic and epigenomic EBF1 alterations modulate TERT expression in gastric cancer. *J. Clin. Invest.* **130**, 3005–3020 (2020).

192. Xu, C. *et al.* Comprehensive molecular phenotyping of ARID1A-deficient gastric cancer reveals pervasive epigenomic reprogramming and therapeutic opportunities. *Gut* **72**, 1651–1663 (2023).
193. Ho, S. W. T. *et al.* Regulatory enhancer profiling of mesenchymal-type gastric cancer reveals subtype-specific epigenomic landscapes and targetable vulnerabilities. *Gut* **72**, 226–241 (2023).
194. Sheng, T. *et al.* Integrative epigenomic and high-throughput functional enhancer profiling reveals determinants of enhancer heterogeneity in gastric cancer. *Genome Med.* **13**, 158 (2021).
195. Kozlova, A. *et al.* Loss of function of OTUD7A in the schizophrenia- associated 15q13.3 deletion impairs synapse development and function in human neurons. *Am. J. Hum. Genet.* **109**, 1500–1519 (2022).

Curriculum Vitae

The Johns Hopkins University School of Medicine

JIN WOO OH April 9th, 2024

EDUCATION

- **Ph.D.** 2024 Biomedical Engineering Johns Hopkins University
  Advisor: Michael A. Beer, Ph.D.
- **M.S.** 2018 Biomedical Engineering Johns Hopkins University
- **B.S.** 2017 Biomedical Engineering Johns Hopkins University
  (Double Major: Applied Mathematics and Statistics)

OTHER TRAINING

- **Consortium trainee** IGVF (2022-present)
  (Impact of Genomic Variation on Function)
- **Consortium trainee** ENCODE (2018-present)
  (Encyclopedia of DNA Elements)

AWARDS AND HONORS

- **Young Investigator's Day Award** (The Mette Strand Research Award)
  2024 Johns Hopkins School of Medicine
- **IGVF Consortium Poster Award**
  2022 IGVF Consortium
- **ENCODE Consortium Team Science Award**
  2022 ENCODE Consortium
- **David T. Yue Memorial Award for Teaching Excellence**
  2017 Johns Hopkins University

TEACHING EXPERIENCE

Teaching assistant at Johns Hopkins University

- Discrete Mathematics 2017
- Statistical Mechanics and Thermodynamics 2016, 2017
- Statistical Physics 2018
- Methods in Nucleic Acid Sequencing Lab 2021
- Biomedical Data Science 2022

PUBLICATIONS

1. **Jin Woo Oh***, David Yao*, Josh Tycko*, Lexi R. Bounds*, Sager J. Gosai*, Lazaros Lataniotis*, Ava Mackay-Smith, Benjamin R. Doughty, Idan Gabdank, Henri Schmidt, Tania Guerrero-Altamirano, Keith Siklenka, Katherine Guo, Alexander D. White, Ingrid Youngworth, Kalina Andreeva, Xingjie Ren, Alejandro Barrera, Yunhai Luo, Galip Gürkan Yardımcı, Ryan Tewhey, Anshul Kundaje, William J. Greenleaf, Pardis C. Sabeti, Christina Leslie, Yuri Pritykin, Jill E. Moore, Michael A. Beer, Charles A. Gersbach, Timothy E. Reddy, Yin Shen, Jesse M. Engreitz, Michael C. Bassik & Steven K. Reilly. "Multicenter integrated analysis of noncoding CRISPRi screens." Nature Methods (2024)
2. **Jin Woo Oh** & Michael A Beer. "Gapped-kmer sequence modeling robustly identifies regulatory vocabularies and distal enhancers conserved between evolutionarily distant mammals." *bioRxiv* (2023; under review)
3. Renhe Luo, Jielin Yan, **Jin Woo Oh**, Wang Xi, Dustin Shigaki, Wilfred Wong, Hyein S Cho, Dylan Murphy, Ronald Cutler, Bess P Rosen, Julian Pulecio, Dapeng Yang, Rachel A Glenn, Tingxu Chen, Qing V Li, Thomas Vierbuchen, Simone Sidoli, Effie Apostolou, Danwei Huangfu, Michael A Beer. "Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions." Nature Genetics. (2023)
4. IGVF. The Impact of Genomic Variation on Function (IGVF) Consortium. *ArXiv* (2023)

* co-first authors

TALKS

1. "Multicenter integrative analysis of noncoding CRISPR screens" (ENCODE consortium meeting, 2022, co-presented)
2. "Sequence alignment using gapped-kmer features identifies conserved orthologous mammalian enhancers with high precision." (ENCODE consortium Meeting, 2021)

POSTERS

1. "Functional characterization of the mammalian genome" (American Society of Human Genetics, 2023**;** co-presented)
2. "Gapped-kmer based machine learning and sequence alignment identify conserved regulatory vocabularies and enhancers in evolutionarily distant mammals" (IGVF consortium Meeting, 2022)
3. "Identifying subsets of cell-specific regulatory sequence features from heterogeneous tissue-specific chromatin state data using predictive gapped k-mer features" (Cold Spring Harbor Laboratory- The Biology of Genomes Conference, 2019)
4. "Identifying subsets of cell-specific regulatory sequence features from heterogeneous tissue-specific chromatin state data using predictive gapped k-mer features" (ENCODE Consortium Meeting, 2019)