Decoding the Non-coding Genome: Novel Technologies for the Characterization of Non-coding Elements and Variation

by

Sierra S. Nishizaki

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Genetics and Genomics)
in The University of Michigan
2020

Doctoral Committee:

Associate Professor Alan Boyle, Chair
Professor Anthony Antonellis
Assistant Professor Kenneth Kwan
Associate Professor Ryan Mills
Associate Professor Elizabeth Speliotes

Sierra S. Nishizaki
ssnishi@umich.edu
ORCID iD: 0000-0002-4288-4567

To my families
The ones who raised me
The ones I've chosen
The ones I've made
You give me purpose

ACKNOWLEDGEMENTS

The work here would not have possible without the combined support of so many wonderful people.

To the Boyle lab for making Michigan feel like home. Thank you for unabashedly being yourselves and for always sharing your enthusiasm for science and life. To my wet-lab family, Jessica, Torrin, and Melissa, you made coming to lab everyday a joy. I cherished our walks to the café, pet play dates, kayaking races, trips up north, and so much more. I consider you all to be lifelong friends.

To my family and friends near and far. Your encouragement keeps me going. To my mother, thank you for inspiring my love of learning. To Becca, Sadie, Paloma, Amy, Bobby, and Courtney I am so excited to see where you go in life. To Pelle, Jacob, An, and Tim thank you for being supportive rubber ducks. To Samantha, for remininding me to love life. To future scientists Xander, Aidan, and Liam, may you always stay curious. To Rose, who has stayed by my side throughout graduate school. Thank you for moving your whole life to Michigan, for listening to all of my practice talks, and for helping me pack pipette tips. You are my corner, and the best wife and mother to our children I could have asked for. To Griffin and Rowan, thank you for learning to sleep through the night at 4 months so that I could write this thesis.

To the amazing folks in the Human Genetics Department and beyond at Michigan, you made this work possible. To Dr. John Moran, Dr. Shigeki Iwase, and Dr. Jeff Kidd, for being generous with your advice and resources. To the Human Genetics staff, Sue Kellogg, Kim White, and Molly Martin for all of their kindness and for holding the department together. To Dr. Michael Boehnke and the GSTP for introducing me to new ways to think about research. To my committee for their encouragement and infectious passion for science. A special thanks to Dr. Anthony Antonellis for his unyielding support, allowing me to use his zebrafish facilities, and for his wisdom.

Finally, to Alan, for being my PhD mentor. Thank you for the honor of being your first graduate student. It has been a privilege to experience the lab grow over these past 6 years, and to watch you mature as a mentor. I deeply appreciate your innovative approach to research, your commitment to your students, and your trust in our abilities as scientists. Thank you for building a community by bringing the lab together, for planning trips to whirly ball and escape rooms, and for welcoming us into your home for lab parties. I will always treasure the time I spent as a member of the Boyle lab.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

One of the key frontiers in genomics research is decoding the function of non-coding sequence and variation. Non-coding sequence, once thought to be junk DNA, is now known to regulate gene expression in a tissue-specific manner, and is frequently found to be mutated in cases of complex human disease. Despite their importance in human disease, non-coding regions are vastly understudied compared to protein coding regions. This is in part due to the abundance of non-coding sequences currently predicted to comprise 98.8% of the genome compared to protein coding regions, which make up only 1.2%. To complicate things further, most of this sequence is non-functional. A non-coding mutation may lead to a change in gene expression or a difference in human phenotype, yet it could show no change in gene expression at all. Therefore, there is considerable demand for novel computational and experimental tools focused on identifying functional non-coding sequences, and prioritizing variation associated with gene expression regulation and human disease.

The focus of the work in this dissertation is the development of novel tools to identify functional non-coding regulatory sequences, and to prioritize the variation that falls within these sequences. I will introduce the following computational tools, the SNP Effect Matrix Pipeline (SEMpl) and the SNP Effect Matrix Pipeline with Methylation (SEMplMe). These methods integrate data from genome-wide annotations of functional elements, such as sites of transcription factor protein binding (ChIP-seq), open chromatin (DNase-seq), and DNA methylation (WGBS), to gener-

ate predictions of the consequences of nucleotide and methylation changes to binding affinity in transcription factor binding sites. As transcription factor binding sites are the building blocks of larger regulatory sequences, such as regulatory elements, functional alterations caused by the introduction of a variant or DNA methylation may lead to aberrant gene expression. SEMpl and SEMplMe are easy to use tools to help researchers prioritize the hundreds of putative regulatory variants that emerge from high-throughput studies, such as genome-wide association studies. This will greatly increase the rate at which non-coding variation can be experimentally validated.

I will also introduce experimental tools focused on identifying larger blocks of regulatory non-coding sequence: cis-regulatory elements. Cis-regulatory elements are sequences that are able to alter or drive gene expression. Currently, a large body of information exists for regulatory elements that are associated with an increase in gene expression, known as positive regulatory elements. However, regulatory elements associated with a decrease in gene expression, also known as negative regulatory elements, are comparatively understudied. To help fill this gap in knowledge between positive and negative regulatory elements, I helped develop two novel methodologies that are able to invert negative regulation into a positive reporter signal. The observed positive output allows negative regulatory elements to be characterized in a spatio-temporal manner *in vivo* in whole animals. This advancement will allow negative regulatory elements to be studied in a manner similar to what has already been achieved for positive regulatory elements for the first time.

Together, the studies in this dissertation investigate non-coding regulatory sequence genome-wide through the development of novel tools which prioritize regulatory variation and identify and characterize regulatory elements.

CHAPTER I

Introduction

1.1 Introduction to the Non-coding Genome

Prior to the completion of the Human Genome Project in 2001, scientists believed that non-genic sequences in the genome represented 'junk DNA' – potentially vestigial sequence no longer playing an important role in human development. However, the results of this seminal study and many studies since have found that there exist only approximately 20,000 genes in the human genome, making up about 1.2% of the total genomic sequence [1]. This surprisingly low number of genes hints that additional mechanisms must be in place for these 20,000 protein-producing units to generate all of the complexity required to form the multitude of tissue-types found in humans. Scientists have now come to appreciate that the other 98.8% of the genome is not simply 'junk', but functions as an additional level of gene expression regulation – altering the availability or rate at which genes are transcribed. These functional, non-coding regulatory sequences can be cell-type specific, allowing for different expression patterns across the same sets of genes to generate distinct cell types. However, while coding regions of the genome have been well annotated, non-coding regulatory regions remain vastly understudied.

Non-coding regulatory sequence is primarily divided into discrete regions known

as cis-regulatory elements. These include promoters - which fall directly upstream of the transcription start site of their target gene and initiate gene expression through the recruitment of transcription machinery, enhancers - which recruit activator proteins to increase target gene expression likely through DNA looping to boost activity at the promoter, silencers - which recruit repressor proteins to decrease target gene expression though a variety of mechanisms, and enhancer blockers - which limit the range of enhancer activity, likely through CTCF-mediated DNA looping into regulatory-activity-insulated topological domains. Regulatory sequences are classically characterized by their activity and are made up of transcription factor binding sites which facilitate protein-DNA interactions. These transcription factor binding sites are where activator or repressor proteins are recruited, and mutations within these sites can lead to transcription factor binding affinity changes, aberrant downstream gene expression, and human disease [2].

1.2 Non-coding Variation is Associated with Human Disease

One of the most surprising results to emerge from genome-wide association studies (GWAS) is that the majority of variation in the human genome associated with disease is found in non-coding regions. Though it has long been known that 95% of genomic variation falls into non-coding sequences, it was unexpected that 88% of potentially disease causing variants identified by GWAS would exist in intronic or intergenic regions [3]. Additionally, simulations have predicted that regulatory regions may account for up to 79% of imputed heritability of examined human traits [4]. In support of this finding, many non-coding variants have been found to contribute to human disease phenotypes [5].

Variants falling into transcription factor binding sites found in regulatory sequence can be categorized by their effect on transcription factor binding (Figure 1.1) [6]. The most common is a change to transcription factor binding affinity. For example, a variant associated with Hirschsprung disease in a putative enhancer was found to decrease reporter gene expression six-fold compared to the wild-type allele, consistent with decreased transcription factor binding at this locus (Figure 1.1A) [7]. Variants strongly associated with Type 2 diabetes and colorectal cancer have been seen to significantly increase reporter gene expression, consistent with increased transcription factor binding (Figure 1.1B) [8]. Ablation of transcription factor binding is also a known consequence of non-coding variation, as was seen in a variant within an AP2 transcription factor binding site associated with cleft lip (Figure 1.1C) [9]. There are other instances where a variant leads to ablation of the binding of its original transcription factor while concurrently gaining a site for an aberrant transcription factor (Figure 1.1E). This was observed in a variant associated with myasthenia gravis in which an SP1 site was disrupted, generating a novel NF-kB transcription factor binding site [10]. Variation in nonfunctional non-coding sequences may also produce novel transcription factor binding sites (Figure 1.1F). This type of variation has been noted in patients with high cholesterol where a novel binding site for the known activator protein C/EBP is produced from a common single nucleotide polymorphism [11]. It has been postulated that variation altering a transcription factor binding site may lead to an additional transcription factor binding alongside the original, however I have yet to identify published evidence of this to date (Figure 1.1D).

Though the majority of validated regulatory elements are promoters and enhancers, silencers and enhancer blockers are predicted to play a major role in human disease. Known disruptions of enhancer blockers have been found to arise from copy

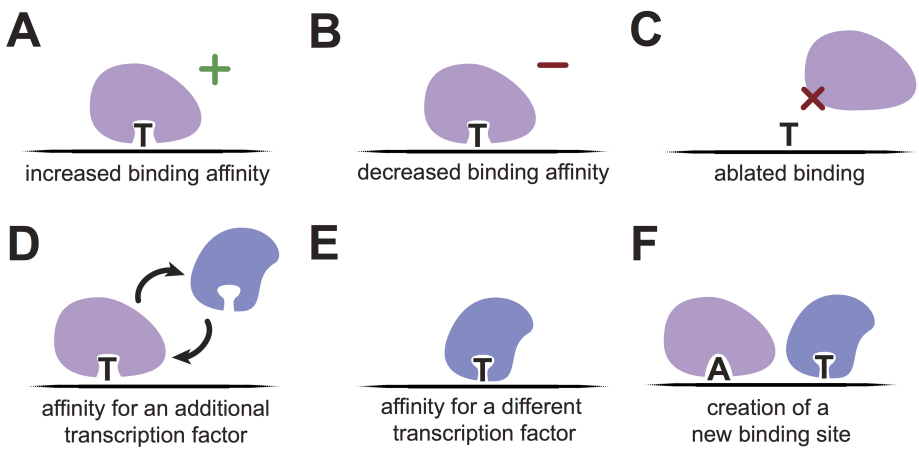


Figure 1.1: Changes to transcription factor binding affinity following the introduction of a variant in a transcription factor binding site. A. A variant leading to increased binding affinity. B. A variant causing decreased binding affinity. C. A variant leading to ablated binding affinity. D. A variant which leads to affinity of an additional transcription factor. E. A variant which ablates binding of the canonical transcription factor, and generates a binding site for a new transcription factor. F. A variant which creates a transcription factor binding site where there was previously none.

number variations or translocated sequences where aberrant enhancer blocker activity disrupts normal regulatory networks. This is best characterized in cases of limb malformation [12]. Additionally, many regulatory elements currently thought to contribute to enhancer activity, may in fact function as silencers. For example, ultraconserved non-coding variants in patients with holoprosencephaly were identified within a putative regulatory sequence which showed no enhancer activity by an *in vivo* reporter assay [13]. Though the researchers speculate this may be the case of a failure in the zebrafish *in vivo* reporter assay to fully recapitulate human regulatory enhancer activity, it can be postulated that this sequence confers silencer activity, which would not be detected by their classical enhancer reporter assay.

1.3 Annotating Regulatory Sequence

While the Human Genome Project set out to sequence all coding sequence within the human genome, the Encyclopedia of DNA Elements (ENCODE) Project is now working to annotate the non-coding human genome [1]. Towards this goal, ENCODE generates high-throughput datasets of features associated with functional DNA (Figure 1.2). This includes chromatin immunoprecipitation followed by sequencing (ChIP-Seq) which measures genome-wide binding of proteins and histone modifications, DNase I hypersensitive sites sequencing (DNase-seq) that measures regions of open chromatin where proteins are more likely to bind and genes are more likely to be expressed, and Whole genome bisulfite sequencing (WGBS) which measures the average amount of DNA methylation at each nucleotide. These data provide a substantial baseline of features associated with functional regulatory sequence. For example, histone modifications are widely used as a proxy for putative regulatory elements, such as promoters and enhancers [14]. Transcription factor protein binding in non-coding sequence may also indicate the presence of regulatory sequence, such as an enhancer or silencer.

Additional annotations also correlate to functional non-coding regions, including sequence conservation from multiple species alignments and catalogs of human variation. Comparative genomics is a tool often used for genic sequence interpretation and can be applied to the non-coding genome to identify regions of sequence undergoing positive selection [16]. These typically include sequences which maintain >70% sequence identity across distantly related species, as with human and Drosophila. Unlike genic sequences, regulatory non-coding sequences are less likely to be conserved, possibly due to the redundant function of regulatory elements [17].

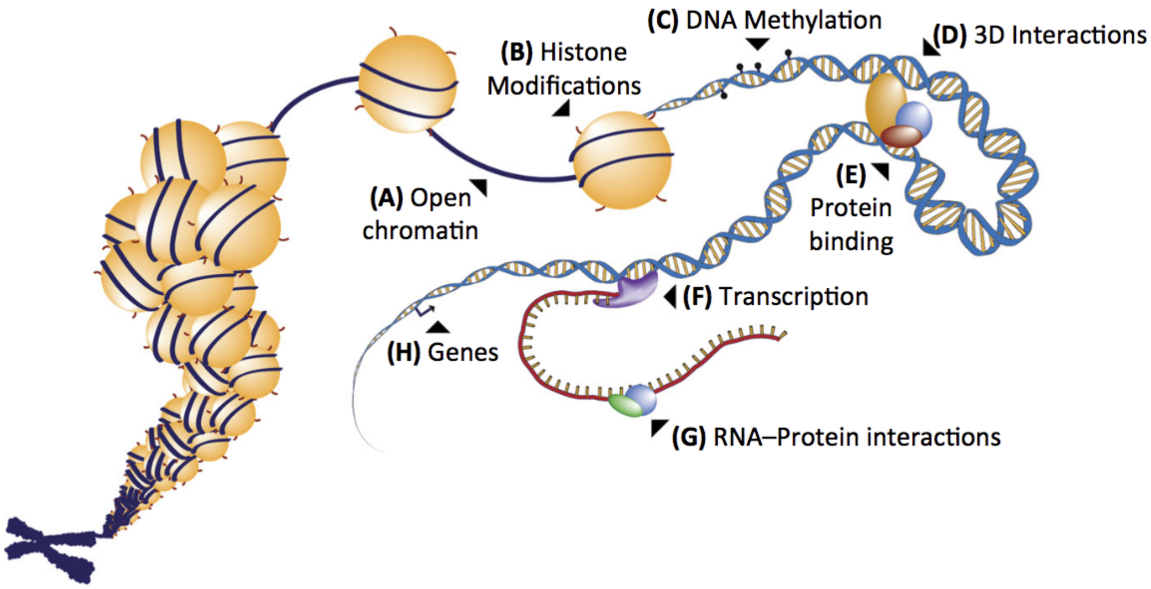


Figure 1.2: Datasets generated by ENCODE. A. Open chromatin from analyses such as DNase-seq, ATAC-seq, and FAIRE-seq. B. Histone modifications from ChIP-seq. C. DNA methylation from whole genome bisulfite sequencing (WGBS). D. 3D interactions from analyses such as Hi-C and ChIP-PET. E. Protein binding from transcription factor ChIP-seq. F. Transcription from RNA-seq. G. RNA-protein interactions from analyses such as icLIP. H. Gene annotations from analyses such as RT-PCR. Figure adapted from Diehl and Boyle [15].

Large-scale sequencing initiatives, such as the 1000 Genomes Project, are utilized to calculate positive and negative selection occurring within the human genome [18]. Additionally, these databases can help to identify discrepancies in the frequency of non-coding variants to those in the general population. The datasets presented here only represent features associated with functional sequence. Further experimental validation is required to verify predicted functional sequence and putatively causal variation.

1.4 Computational Predictions of Regulatory Sequence

While experimental validation is the gold standard for confirming functional non-coding sequences, it can be difficult to prioritize which of these regions may be

functional out of the >2.9 billion base pairs in the human genome. Focusing on transcription factor binding sites, which are more likely to have regulatory activity, can narrow this down to 240 million base pairs. However, as genomic regions are traditionally tested for functional regulatory activity <1000bp at a time, validating non-coding variation remains a herculean task. Additional methodologies are required to further prioritize putatively functional non-coding genomic sequences in order to make the experimental validation of non-coding regions more accessible.

Computational non-coding annotation tools use genomic features associated with functional sequence to predict how likely a genomic region is to have functional or regulatory activity. These tools utilize many of the same types of high-throughput datasets generated by ENCODE, such as ChIP-seq and DNase-seq, as well as additional annotations, including sequence conservation derived from software such as phyloP, and 3D chromatin contacts from assays like Hi-C and ChIA-pet [1, 19, 20, 21]. Analyses that include direct association tests can be utilized to identify expression quantitative trait loci (eQTL), which are markers of genetic variation directly associated with gene expression changes [22]. Together, these annotations are used by non-coding SNP annotation tools heuristically or to train machine learning models to generate quantitative scores predicting the likelihood that variants are associated with, but not necessarily causal of, gene expression changes [22, 23].

Tools like RegulomeDB previously used a heuristic scoring system of ChIP-seq, DNase-seq, and eQTL data to generate prediction scores [24]. However, the newer version of RegulomeDB (RegulomeDB 2.0) utilizes these annotations, as well as scores from another non-coding SNP annotation method, DeepSEA, in a random forest machine learning model to better prioritize non-coding variation [25, 26]. Additional tools, such as CADD and deltaSVM, have also emerged as popular choices

of machine learning methods for predicting functional non-coding variations, and for coding variations in the case of CADD [27, 28]. Despite the benefits these tools provide, <40% of GWAS failed to follow up on variation falling into non-coding regions using one of these tools [23]. This is due to the need for more robust predictions of non-coding variation that can arise from increasing the types of function-associated annotations used in these methods. The high cost of experimental follow-up, paired with a lack of tools to validate and characterize non-coding variation and regulatory regions is also a large contributor.

1.5 Experimental Validation of Regulatory Sequence

In addition to generating genomic annotations of features associated with functional regulatory sequence, such as those available from ENCODE, experimental assays have been developed to validate the function of putative regulatory elements and variants.

One such assay is the classical reporter assay that is used to validate and characterize promoter and enhancer activity in cells or whole organisms. Enhancer assays function by inserting a putative enhancer upstream of a minimal promoter driving a reporter gene, such as GFP, on a plasmid backbone. In cases where the tested sequence has enhancer activity, GFP expression will increase above minimal promoter expression levels. Similarly, promoter assays involve insertion of putative promoters into a plasmid backbone directly upstream of GFP. Observable expression is predicted if the putative promoter sequence is able to promote GFP expression. This type of assay can be used to generate transgenic animal models to examine spatio-temporal activity of regulatory elements, and is commonly employed in mice using

LacZ as a reporter, and in zebrafish using a GFP reporter [29, 30]. Large-scale initiatives, such as the VISTA Enhancer Browser, are working to characterize the spatio-temporal activity of an abundance of regulatory elements using reporter assays [31].

The classic reporter assay has been extended for use in high-throughput studies of enhancer activity. Massively parallel reporter assays (MPRAs) use the logic of classical reporter assays, but are able to investigate whole sequencing libraries in parallel by labeling plasmids with unique barcodes in place of a reporter gene [32]. Following transfection into a cell line, the expressed barcodes are purified from mRNA and sequenced alongside the DNA plasmid library to link the tested enhancer with its corresponding barcode. Enhancers matching an expressed barcode are considered active. A similar method named self-transcribing active regulatory region sequencing (STARR-seq) is also able to evaluate entire sequencing libraries in parallel [33]. This assay circumvents the use of a barcode by inserting putative enhancers downstream of a minimal promoter to drive their own expression. Any enhancer recovered after mRNA sequencing from transfected cells is considered active in the tested cell-type.

Classical reporter assays have also been adapted to assess the regulatory activity of putative silencers and enhancer blockers [34]. These assays rely on placing putative silencers upstream of a known enhancer and minimal promoter that drives reporter gene activity, with the output being a reduction in expression. A similar assay exists for enhancer blockers wherein a putative enhancer blocker is inserted between the enhancer and minimal promoter. This expected reduction in reporter signal greatly increases the number of false positives yielded by these assays, and makes them impractical for use *in vivo*. These disparities in reporter assay technology between promoters/enhancers and silencers/enhancer blockers is a key driver of the imbalance

in studies between these regulatory element types.

In addition to reporter assays, the function of a regulatory variant can be evaluated by assessing the transcription factor protein-DNA binding differences between variant and wild-type alleles [9]. This is commonly achieved using electrophoretic mobility shift assay (EMSA) analysis, in which transcription factors and their potential binding sites are mixed tougher and allowed to form a complex *in vitro*. These mixtures are run through an agarose gel where protein-DNA complexes appear shifted compared to unbound DNA. The shifted bands can be quantified to determine relative binding affinity between DNA fragments of different alleles. Alternatively, mass-spectrometry has been used to quantify the abundance of transcription factor binding to risk and non-risk alleles from heterozygous patient cell cultures [35]. Relative binding of alleles can be measured using this method by examining the ratios of allelic peaks as a proxy for allelic load following immunoprecipitation of a transcription factor of interest.

Novel technologies, such as clustered interspaced short palindromic repeats (CRISPR), have recently been employed to directly study patient mutations. For example, CRISPR-mediated genomic rearrangements matching patient genotypes were used to study human limb malformations [12]. This method found that disrupted enhancer blocker activity within topologically associated domain boundary regions lead to aberrant gene expression contributing to patient syndactyly, brachydactyly, and F-syndrome. Generating patient mutations in regulatory sequence followed by assessment of global gene expression levels by RNA-seq can determine if a non-coding variant is functional, as well as delineate genes in its regulatory pathway. Similarly, novel chromatin conformation technologies, like Hi-C and ChIA-PET, can link validated regulatory elements to their target gene, further unraveling the regulatory

landscape of the genome [19].

1.6 Overview of Dissertation

Recent high-throughput analyses have established that the majority of variation in the human genome associated with disease falls in non-coding regions. The bulk of this variation is likely found in regulatory sequence - regions of DNA able to alter tissue-specific gene expression. However, much of the regulatory sequence predicted to be in the genome has yet to be identified, and it remains an undertaking to predict the possible consequences of variation and epigenetic changes falling within this sequence.

In this dissertation I describe novel tools to interpret non-coding variation and methylation, as well as validate and characterize non-coding regulatory regions. Chapter 2 examines current computational tools to prioritize and predict the function of non-coding variants. Chapter 3 introduces a novel tool, SEMpl, to help prioritize non-coding variation by predicting the effect of variation falling into transcription factor binding sites. Chapter 4 extends the SEMpl method to include predictions of the effect of DNA methylation on transcription factor binding. Chapter 5 establishes a widely-used experimental technology, the *lac* operator-repressor system, as functional in zebrafish cells. Chapter 6 utilizes this *lac* operator-repressor system, as well as CRISPR technologies, to generate novel inversion assays able to validate and characterize the spatio-temporal activity of silencers and enhancer blockers in zebrafish. Finally, chapter 7 concludes with perspectives on how this research can contribute to the field, and future directions of the study of non-coding variation going forward.

CHAPTER II

Mining the Unknown: Assigning Function to Non-coding SNPs

2.1 Abstract

One of the formative goals of genetics research is to understand how genetic variation leads to phenotypic differences and human disease. Genome-wide association studies (GWASs) bring us closer to this goal by linking variation with disease faster than ever before. Despite this, GWASs alone are unable to pinpoint disease-causing single nucleotide polymorphisms (SNPs). Non-coding SNPs, which represent the majority of GWAS SNPs, present a particular challenge. To address this challenge, an array of computational tools designed to prioritize and predict the function of non-coding GWAS SNPs have been developed. However, fewer than 40% of GWAS publications from 2015 utilized these tools. We discuss several leading methods for annotating non-coding variants and how they can be integrated into research pipelines in hopes that they will be broadly applied in future GWAS analyses.

2.2 Toward the Goal of Understanding Variation

Genome-wide association studies (GWASs) are a popular method of linking genomic variation with human disease and have produced over 100,000 genomic re-

gion–disease associations to date [36]. These studies are successful at narrowing down potential variants associated with a disease; however, they are incapable of determining causative single nucleotide polymorphisms (SNPs) on their own. GWAS variants are typically screened using a set of lead SNPs, which are informative but often not causative. The causative SNP may lie anywhere within the linkage disequilibrium (LD) block surrounding the lead SNP, but these can span over 100 kb and often contain over 1000 individual SNPs. Improvements in the identification of causative SNPs from GWASs will advance our understanding of disease mechanisms and reveal potential therapy targets. Fine mapping techniques using high-throughput imputation have the potential to refine GWAS SNPs in LD loci down to a testable number, and can be used to make predictions of SNP associations with a phenotype when paired with statistical predictions of association, such as Bayesian refinement [8, 37]. Indeed, combining fine mapping and functional annotations has yielded important discoveries. For example, Bauer et al. identified a single variant in LD with a GWAS locus associated with hemoglobin disorders, which disrupts the motif of an enhancer in a regulator of fetal hemoglobin, BCL11A, and now represents an attractive therapeutic target for the treatment of hemoglobinopathies [38]. However, this methodology requires dense genotyping and large sample sizes, and may not be effective for all loci. Because of these challenges, researchers have now developed many computational tools designed to assist with the prioritization of GWAS SNPs to reduce the resources and time needed to experimentally validate causative SNPs [39, 40]. Although the vast majority of GWAS-implicated SNPs are found in non-coding sequence, the majority of SNP annotation tools only annotate SNPs in coding regions of the genome [3]. This is in part because non-coding SNPs are more challenging to annotate than SNPs in coding regions where the consequences of variation

are better understood. Landmark initiatives now provide sufficient data to begin the task of predicting and prioritizing functional SNPs in non-coding DNA. These include catalogs of human variation (1000 Genomes Project, International HapMap Project), annotations of functional elements [Encyclopedia of DNA elements (ENCODE)], and conservation information derived from multiple species alignments [18, 1]. Since 2010, a handful of tools to annotate non-coding SNPs have been released. These tools provide hypotheses to the functional nature of non-coding SNPs, a powerful first step that reduces the pool of possible variants for experimental follow-up. However, many studies do not take advantage of these tools. In fact, of 44 GWASs released in 2015, only 16 use any sort of non-coding SNP annotations for variant follow-up (see Table S1 in the supplemental information online). Regulatory variants can have dramatic effects on gene regulation. Kasowski et al. [2, 41] initially demonstrated this on a genome-wide scale by showing allele-specific binding of the transcription factor (TF) nuclear factor-kappa B and CCCTC-binding factor (CTCF). Subsequently, Degner et al. [42] demonstrated that a single variant can result in both disruption of TF binding and alteration of chromatin accessibility. Other studies have demonstrated similar dramatic effects of non-coding variation on regulatory networks and gene expression control mechanisms [43, 44]. These findings suggest a mechanistic link between regulatory variation and disease phenotypes. It is clear that, by restricting experimental follow-up to easily classified variants, we likely miss a substantial proportion of variants directly relevant to disease. Through broad application of non-coding SNP annotation tools to GWAS, we can improve our understanding of genetic disease predispositions. In the following sections, we review several leading non-coding SNP annotation tools, examine their strengths and limitations, and discuss how they can be integrated into GWAS pipelines to augment their findings.

Their incorporation will significantly accelerate discovery of disease-causal variants from GWASs and provide vital information to shape hypotheses about their function.

2.3 Annotation of Functional SNPs

Tools for SNP annotation can take advantage of diverse genomic data types to provide putative functional annotations or predict functional effects. Here, we divide the tools into three categories: functional, conservation, and machine learning based. While all of the tools reviewed here utilize functional data, conservation-based tools also include measures of conservation, and machine learning tools may incorporate multiple lines of evidence, including functional annotations and conservation.

Functional Annotation

GWAS SNPs have been shown to be enriched for functional annotations, with 81% of GWAS LD regions containing at least one functional SNP [45]. Many types of high-throughput assays are used to predict features associated with putative regulatory function in the non-coding genome, including DNase I hypersensitive sites sequencing (DNase-seq), assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq), formaldehyde-assisted isolation of regulatory elements-sequencing (FAIRE-seq), TF ChIP-seq, histone modification ChIP-seq, and expression quantitative trait loci (eQTL) analysis (Figure 2.1) [46]. In some cases, genomic features such as distance from the nearest gene, guanine–cytosine content, predicted TF binding motifs, and manual annotations of published variants are also included. As many of these analyses identify cell-type specific interactions, the range of conditions (cell-types, stages of development, etc.) for which data are available

restricts the range of functional elements a tool is able to detect. In addition, these methods will miss functional elements that do not coincide with known annotation co-occurrence patterns. Examples of tools that annotate non-coding SNPs using only functional genomics information include the Ensembl Variant Effect Predictor (VEP), RegulomeDB, and Functional Identification of SNPs (FunciSNP) [24, 47, 48].

Conservation

In addition to functional genomics data, including conservation data allows variants to be ranked based on well-accepted measures of evolutionary constraint. Conservation is typically determined by multiple sequence alignments, from which we can estimate rates at which different categories of genomic regions have evolved over time. Conservation can be measured by comparing the substitution rate within a genomic region of interest to an estimate of the neutral substitution rate. Regions with a significantly lower-than-expected substitution rate are considered to be conserved, and are therefore likely under functional constraint. However, because humans have likely undergone recent rapid adaptations in tissue-specific regulation, strict conservation-based approaches to regulatory element detection may miss critical human-only advancements in tissue types such as the brain [49]. Methods that integrate conservation into their annotation include ANNOVAR, HaploReg, GWAS3D, and fitCons [50, 51, 52, 53].

Machine Learning

Machine learning algorithms have recently become popular for SNP annotation because of their multifaceted predictions based on robust statistical methods. These powerful tools are able to build complex predictive models of SNP function [54]. All

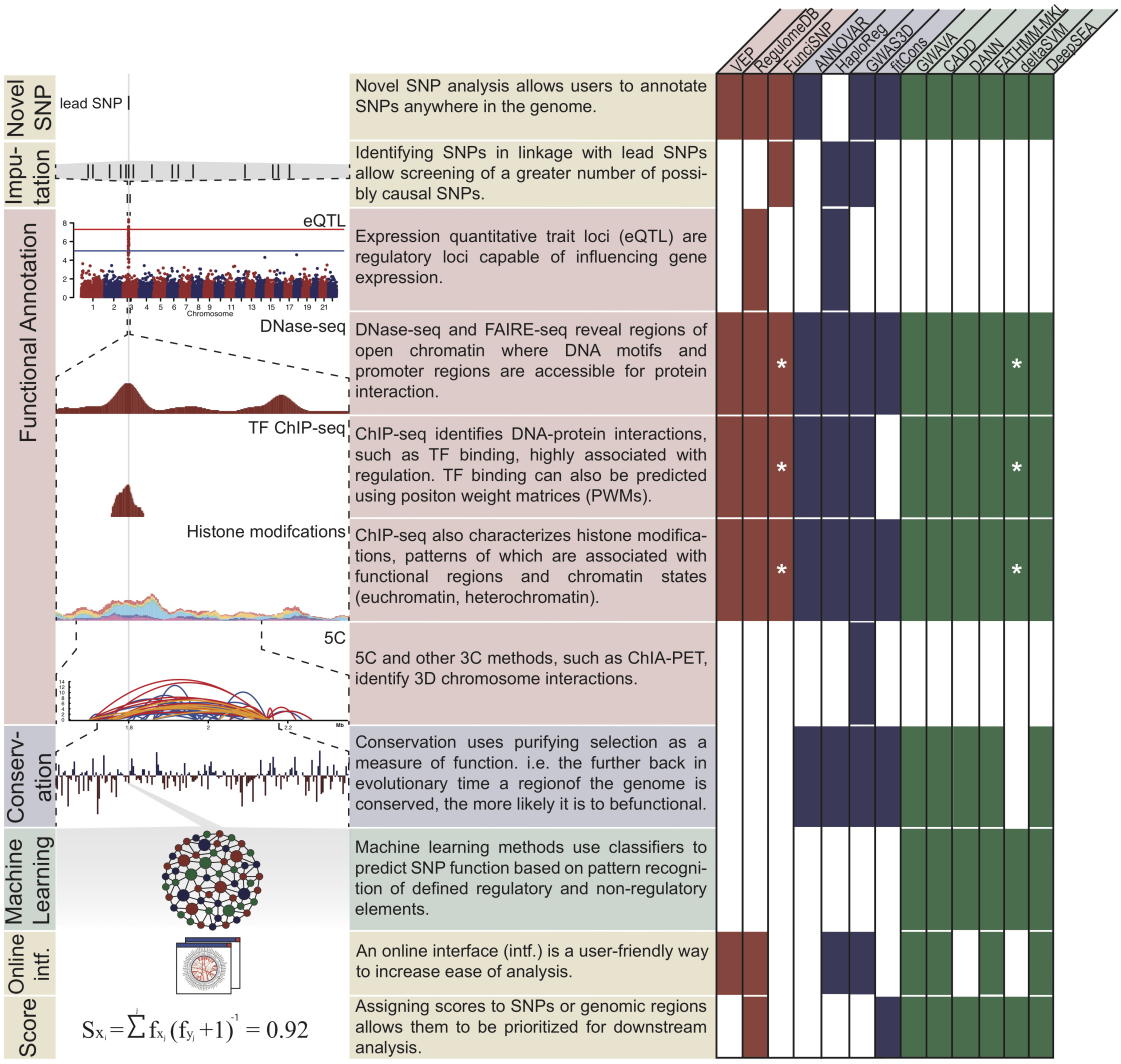


Figure 2.1: Data and tools used to analyze non-coding variants. Single nucleotide polymorphism (SNP) aligned with functional (red) and conservation (blue) data, machine learning methods (green), and tool features (yellow). Each tool discussed in this perspective is labeled with annotation types used in its non-coding variant analysis platform. * represents optional input data sets supplied by the user. Abbreviations: 3C, chromosome conformation capture; 5C, chromosome conformation capture carbon copy; CADD, combined annotation-dependent depletion; ChIA-PET, chromatin interaction analysis by paired-end tag sequencing; DANN, deleterious annotation of genetic variants using neural networks; DNase-seq, DNase I hypersensitive sites sequencing; eQTL, expression quantitative trait loci; FAIRE, formaldehyde-assisted isolation of regulatory elements; FunciSNP, Functional Identification of SNPs; GWAVA, genome-wide annotation of variants; TF, transcription factor; VEP, Variant Effect Predictor.

these methods incorporate functional data and most incorporate conservation data to train their prediction models, each using different models and approaches. Though powerful, machine learning methods are susceptible to biases found in training sets and annotations such as enrichments of variants near genes, gaps in functional annotations, or overfitting due to suboptimal parameterization or insufficient training data. Much care is required to limit the effect these biases have on pattern prediction [27, 55]. In addition, the basis for functional categorization may not be intuitive, as reasons for annotation may not be directly reported in the results. Current methods using machine learning to prioritize candidate functional variants include genome-wide annotation of variants (GWAVA), combined annotation-dependent depletion (CADD), deleterious annotation of genetic variants using neural networks (DANN), FATHMM-MKL, deltaSVM, and DeepSEA [27, 28, 56, 40, 57, 58]. Importantly, the data sets used to train machine learning methods can alter which variants they call. As there is currently no gold-standard training set for detrimental non-coding variants, non-coding annotation tools use a variety of data sets to train their algorithms. For example, GWAVA and FATHMM-MKL use manually curated disease-associated variants from the Human Gene Mutation Database [59], a data set composed of experimentally validated and likely disease-associated variants. However, these databases do not contain randomly sampled variants from across the genome and so are subject to ascertainment bias. GWAVA and FATHMM-MKL attempt to mitigate these biases by sampling nearby nondisease-associated variants. By contrast, CADD and DANN use randomly simulated deleterious variants and conservation between humans and chimp to generate hypothetical sets of deleterious and nondeleterious variants. Though this approach may reduce selection bias, using randomly simulated variants risks capturing nondeleterious alleles in the deleterious

training set, and deleterious alleles in the nondeleterious training set, as it does not use any experimental measure of deleteriousness. Newer methods such as deltaSVM and DeepSEA are forgoing the generation of detrimental SNP training sets altogether in favor of strict functional annotation to identify cell-specific regulatory elements and randomly sampled matched control regions. However, similar to randomly simulated variation, there is no guarantee that randomly selected control regions do not confer some regulatory function or undiscovered disease association. Finally, as machine learning methods do not provide functional annotations alongside predictions of SNP deleteriousness, additional analysis using a functional- or conservation-based tool or manual functional annotation may still be needed to suggest hypotheses for how functional SNPs affect their associated disease phenotypes.

2.4 Integrating SNP Annotation into the GWAS Pipeline

The aforementioned tools offer a powerful way to improve the resolution of GWAS. By integrating them into GWAS pipelines, as shown in Figure 2.2, a list of SNPs in LD with the lead SNP can be annotated and ranked according to their likelihood of function.

Many of these tools approach the variant annotation from the perspective of an individual variant rather than considering all variants in LD with the reported SNP (e.g., RegulomeDB, CADD/ DANN, deltaSVM). For these methods, preprocessing with tools, such as IMPUTE2, is necessary to identify SNPs in linkage with the lead SNP [60]. Other SNP annotation tools (e.g., FunciSNP, HaploReg, GWAS3D) incorporate LD SNPs without the need of additional tools. However, it is important for researchers to consider the genomic background of their samples when using tools

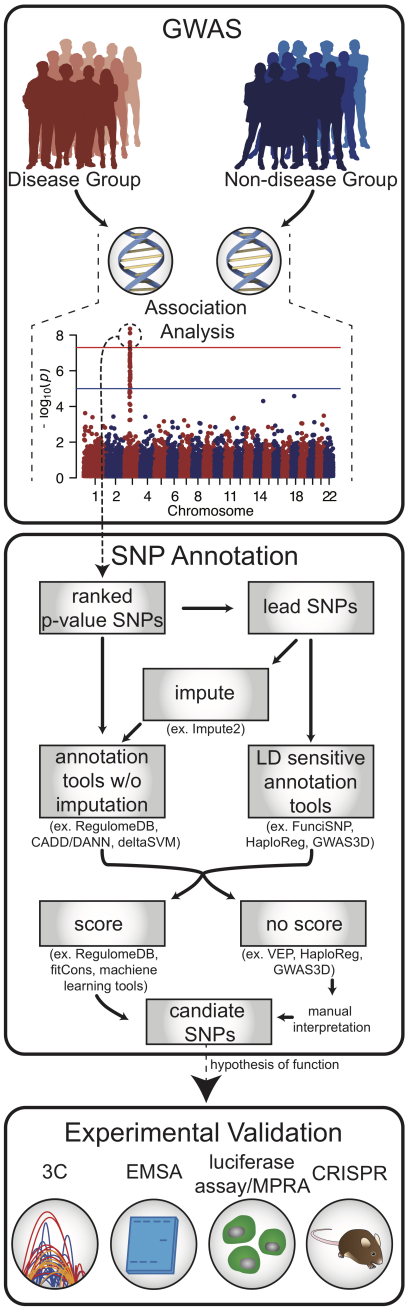


Figure 2.2: Following GWAS analysis, lead SNPs implicated as important in disease risk can be passed to an SNP annotation tool. Annotation tools sensitive to linkage disequilibrium (LD) regions, or who make predictions covering genomic regions can be used directly, while those tools without imputation methods must first be put through an imputation program to make predictions for all SNPs in a region of LD. Once a SNP annotation tool has been implemented, the resulting scores or functional annotations can be used to prioritize candidate SNPs for further experimental validation following generation of a hypothesis of function. Abbreviations: 3C, chromosome conformation capture; CADD, combined annotation-dependent depletion; CRISPR, clustered regularly interspaced short palindromic repeats; DANN, deleterious annotation of genetic variants using neural networks; EMSA, electrophoretic mobility shift assay; FunciSNP, Functional Identification of SNPs; MPRA, massively parallel reporter assay; VEP, Variant Effect Predictor.

that incorporate LD, as regions of LD vary between ethnicities. Because of this, it may be advisable to perform independent imputation as is standard in GWAS before applying these tools. Not only can functional annotations narrow the pool of candidates for experimental follow-up, but also the content of the functional associations (overlapping TF binding sites, chromatin marks, etc.) can suggest casual mechanisms and help direct the strategies used for experimental validation. Annotation tools that provide quantitative scores [such as RegulomeDB, fitCons, and machine learning methods (Figure 2.1)] are particularly well suited to this application. The scores provide a way to directly rank individual SNPs and prioritize them for follow-up. Incorporating expert domain knowledge of the system(s) involved can further guide this process. The associated annotations can provide direct clues to the function of the sequence harboring a SNP of interest, leading to testable hypotheses regarding the tissues, cell types, pathways, target genes, and specific regulatory mechanisms potentially disrupted by a given variant. A final consideration in the use of these tools is the application interface provided to the researcher. Some tools (e.g., FunciSNP, ANNOVAR, deltaSVM) only provide a command-line interface that, while not particularly user friendly, is ideal for integration into bioinformatic pipelines. Conversely, some tools provide Web interfaces with associated graphics and sorting capabilities to allow a noncomputationally focused researcher to perform these analyses with ease, allowing online visualization or opportunities to download scores for further analysis. However, these methods may be difficult to incorporate into automated analysis pipelines. The ideal interface will likely be defined by the research process of each group and should be considered on a case-by-case basis. A recent example of the successful integration of annotation analysis into the GWAS pipeline comes from Higgins et al. [61]. This study examined 31 putative causal SNPs

associated with psychotropic drug response, narrowed down from 2,024 SNPs aggregated across 26 GWAS in the National Human Genome Research Institute (NHGRI) GWAS catalog [36]. They first imputed lead GWAS SNPs from using LD data from HaploReg. Imputed SNPs were then analyzed by SNP annotation methods and additional functional features, including RegulomeDB, HaploReg, and chromatin state. This allowed the authors to identify putative functional SNPs within their LD regions, as well as assign possible regulatory activity to the regions associated with these SNPs (promoter, enhancer, transcribed domain). Finally, by incorporating 3D chromatin interaction data, including GWAS3D analysis, the authors were able to predict cis-regulatory interactions. In total, these predictions provided hypotheses of SNP regulatory activity and interactions, which allow a higher confidence starting point for experimental verification. Another clear demonstration of the power of SNP annotation of GWAS and subsequent experimental validation was recently published by He and colleagues [62]. This study used HaploReg functional annotations, along with known TF binding and histone modification data, to identify multiple novel functional regions and four variants likely to be functional in papillary thyroid cancer. They determined that these variants lead to increased enhancer activity by luciferase assay and increased TF binding by ChIP assay. Using chromosome conformation capture (3C), the authors also identified the gene targets of these enhancers. In this study, the use of non-coding SNP annotation tools, along with additional functional annotations, allowed the authors to distinguish novel enhancers, within which they were able to prioritize and validate SNPs of interest.

2.5 Validation of Tools on Liver SNPs

To demonstrate the applicability and accuracy of these methods, we used four non-coding annotation tools to examine human liver enhancer SNPs previously shown to affect enhancer activity by massively parallel report assay [63](see the 'Methods' section). None of these variants occurs in dbSNP and would be considered de novo variants. The chosen liver data sets were not used to train any of the machine learning methods examined here. Using the default settings for the online interfaces of RegulomeDB, CADD, FATHMM-MKL, and DeepSEA, we found that four of the top seven SNPs that correlated with the greatest change in translational activity were called putatively detrimental or functional by all of the assayed methods, two additional SNPs were called by three of the four methods, and six more SNPs were called by just two of the methods. Only four of the ten SNPs with no effect on transcriptional activity were called benign or nonfunctional by all of the assayed methods. DeepSEA scores had the greatest correlation with the absolute log2-fold change on transcriptional activity ($R^2 = 0.307$), followed by RegulomeDB ($R^2 = 0.262$), CADD ($R^2 = 0.187$), and FATHMM-MKL ($R^2 = 0.168$; Figure 2.3). DeepSEA and RegulomeDB also had the highest agreement when comparing scores ($R^2 = 0.677$). Interestingly,all methods were biased toward predicting SNPs, leading to a decrease in transcriptional activity rather than an increase. Though this trend makes sense for annotation-based methods such as RegulomeDB, the ability to identify SNPs with a positive effect of transcription is surprisingly low for sequence based methods CADD and FATHMM-MKL. These results are striking, as none of the tools examined includes measures of SNP effects on gene expression in their prediction models. Discrepancies among scores given to the same variant by different annotation tools

are not surprising. McCarthy et al. [64] explored the effect of annotation tools on coding variant prediction using ANNOVAR and VEP, and found only an 87% agreement between annotation calls. We expect to find far more discrepancies in non-coding regions of the genome, where markers of regulatory activity are far less understood. As an independent large-scale comparison of these methods has yet to be published, it remains unclear which tool, if any, is generally the most effective.

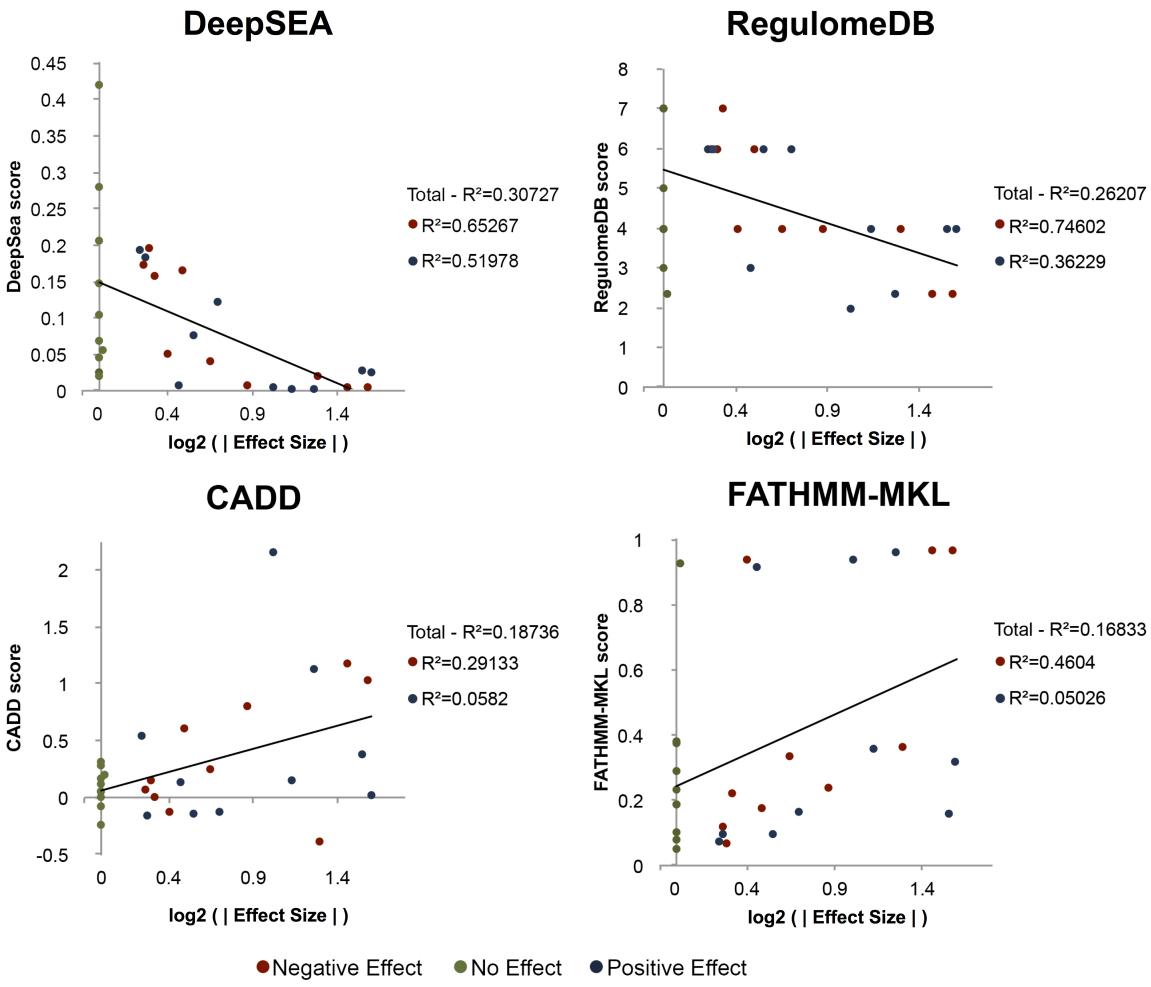


Figure 2.3: Effect size represents the absolute value log2 change of the transcriptional activity of the variant compared with wild type. R2 values are given for all data points, and for positive and negative data points individually. Red points represent variants correlated with a negative effect on transcription, blue points are those correlated with a positive effect on transcription, and green points are correlated with no effect on transcription. Abbreviation: CADD, combined annotation-dependent depletion.

Our comparison demonstrates disparities in the agreement between the calls of

different tools, suggesting that the use of multiple tools in tandem may increase the confidence of called SNPs, and this strategy has been successful in multiple published studies [65, 66, 67]. For example, Chen and colleagues [68] utilized VEP, RegulomeDB, ANNOVAR, and HaploReg to predict the likelihood of function and regulatory feature type for 9,184 non-coding variants from the NHGRI database. Strikingly, they were able to predict regulatory functions for 96% of these variants. Furthermore, they randomly selected three variants from their list for functional testing in a reporter assay, and found all three to have enhancer or silencer activity. These results highlight the promise of using multiple non-coding annotation methods to increase the confidence of predicted casual SNPs. Indeed, combining multiple annotation tools may balance out the biases inherent to single tools, thus yielding more reliable predictions.

2.6 Experimental SNP Validation

Many experimental methods exist for investigating the effects of SNPs, but without specific functional hypotheses, choosing an appropriate method of experimental follow-up is challenging [69, 70]. Integrating annotation tools that report functional information into the GWAS pipeline provides multiple lines of evidence to suggest appropriate tests, including the correct tissue or cell type, how a SNP affects regulation at a locus (e.g., by altering TF binding), the gene target of the regulatory region, and the expression-level effect on the target gene. In particular, functional annotations provided by many tools can suggest the cell type in which a SNP may have an effect. This is particularly crucial in non-coding regions of the genome, as most regulatory regions are tissue specific. Thus, cell-type predictions can inform de-

cisions on which cells to use for *in vitro* analyses, such as luciferase reporter assays, and which tissues to examine in *in vivo* analyses, such as immunohistochemistry. One method commonly used to investigate the disruption of protein–DNA interactions by regulatory SNPs is electrophoretic mobility shift assay. This assay can be used to determine if a protein is capable of interacting *in vitro* with a DNA sequence of interest, and can be used to assay if DNA–protein interactions are perturbed by introducing a SNP [71]. Proteome-wide analysis of SNPs can also be used to identify SNPs producing differential TF binding [72]. Many GWAS findings are distal to any obvious target gene and many regulatory elements have been shown to act on a gene other than the nearest gene [73]. To identify the target for a regulatory region, one can use a 3D genomic assay such as 3C, chromosome conformation capture-on-chip (4C), chromosome conformation capture carbon copy (5C), chromatin interaction analysis by paired-end tag sequencing, Hi-C, Capture C, or Capture Hi-C (CHi-C) [74, 75, 20, 76, 77]. The GWAS3D annotation tool includes a set of 3D interaction data in its annotations and some tools include eQTL information that may give an idea of the gene regulatory interaction. In cases where there are no current data, an assay such as 4C will allow interrogation of all interactions with the significant locus. Following target gene identification, expression changes can be assayed using reverse transcription PCR. However, though these methods can demonstrate regulatory interactions between non-coding sequences and target genes, they cannot discern specific functional effects. Reporter assays offer a complementary approach to the aforementioned methods, offering the ability to directly measure the functional effect of a variant on gene expression levels. They work by placing a regulatory element upstream of a minimal promoter and a reporter gene in a plasmid, which can be transfected into an organism and analyzed for regulatory activity [62]. High-

throughput forms of these assays can be used to measure functional consequences of variation more broadly [33, 32]. Likewise, transgenic animal models, including mice and zebrafish, offer powerful tools to assay the phenotypic effect of mutations *in vivo* [78, 29]. With the discovery of clustered regularly interspaced short palindromic repeats (CRISPR) editing, non-coding variants and structural changes may now more easily be investigated in these more complex model systems [79].

2.7 Concluding Remarks

We believe computational SNP annotation tools will prove invaluable to the interpretation of GWAS SNPs. The tools reviewed here provide annotations and predictions of the regulatory effects of these often-difficult-to-interpret variants using three primary methodologies: functional annotations, conservation, and machine learning (Table 1). Though the majority of GWAS analyses using these methods stop at selecting possible functional variants within an LD region, the power of these annotation methods will come from increasing the speed and ease of experimentally validating putative causal SNPs associated with disease [80]. This improvement will be primarily through reducing the set of variants for experimental follow-up and guiding hypothesis generation regarding their target tissues and regulatory impacts. Validated causal SNPs can then feed back into future development efforts, further refining these techniques and improving their utility. As gaps in functional data are filled and high-throughput sequencing technologies improve, SNP annotation methods will become more powerful. Notably, increased adoption of whole-genome sequencing technology, along with improvements to the technologies themselves, will vastly improve the breadth and resolution of available sequence data. This will allow

not only non-coding variants to be detected, but also improved annotations of structural variation such as copy number variation [81]. In addition, the development of ensemble predictors, similar to those available for coding annotations, would allow users to run several annotation models in parallel, providing the same benefit as implementing multiple tools. The expansion of functional data sets across a wide range of cell types will be key to improving variant predictions for tissue-specific phenotypes. Finally, incorporation of 3D structural data will likely improve our ability to assign regulatory SNPs to their target genes, with additional improvements in our ability to discern their functions and place them in their biological context, a necessary step for critical pharmacogenetic advancements.

The widespread use of non-coding SNP annotation methods will help us predict the effects of genomic variation, elucidate mechanisms and pathways of disease, and bring us closer to understanding the full complexity of the human genome.

2.8 Methods

SNP Selection

Variants in Figure 2.3 were chosen from two human enhancer loci previously examined at a nucleotide level by massively parallel reporter assay [63]. Enhancer loci were divided into fifths, with three SNPs chosen from each region (ALDOB, hg19:chr9:104195570–104195820; ECR11, hg19:chr2:169939182–169939682). From each fifth we selected the two SNPs that correlated with the greatest positive or negative change in transcriptional activity, and the first occurrence of a SNP leading to no change (or in the absence of any such SNP, the variant closest to 0 translational activity) in the region, for a total of 30 SNPs.

Tool	Description	Ref
VEP	VEP incorporates annotations from the Ensembl database, allowing it to make predictions genome-wide as well as predict tissue-specific activity for 13 human cell lines.	McLaren et al. [48]
RegulomeDB	RegulomeDB uses a heuristic scoring system to catalog the likelihood that a given SNP or indel resides in a functional region, using functional data from over 100 cell types.	Boyle et al. [24]
FunciSNP	FunciSNP is an R/Bioconductor package that employs user input annotations to prioritize SNPs, allowing users to customize their annotations to query a cell type of interest.	Coetzee et al. [47]
ANNOVAR	ANNOVAR is a command line tool that uses region-based annotations to annotate non-coding variants and insertions and deletions (indels), in addition to comparing them to known variation databases.	Wang et al. [52]
HaploReg	HaploReg is a searchable repository for SNPs and indels from the 1000 Genomes Project, providing a summary of known annotations for variants within an LD block.	Ward and Kellis [53]
GWAS3D	GWAS3D evaluates SNPs and indels by analyzing their 3D chromosomal interactions and disruptions to TF binding affinity. It outputs scores as well as a circle plot mapping local 3D interactions.	Li et al. [51]
fitCons	fitCons uses the INSIGHT method to predict the probability that SNPs will influence fitness by screening for signatures positive and negative selection using data from three cell types.	Gulko et al. [50]
GWAVA	GWAVA trains on a random forest algorithm using disease mutations from HGMD and control variants from the 1000 genomes project to predict if queried variants are functional.	Ritchie et al. [40]
CADD	CADD trains on a linear kernel support vector matrix using simulated variants as deleterious variants and alleles fixed between human and chimpanzee as control variants.	Kircher et al. [27]
DANN	DANN trains on a nonlinear learning neural network algorithm using the same training set data (fixed alleles vs. simulated variants) as CADD.	Quang et al. [56]
FATHMM-MKL	FATHMM-MKL implements a kernel-based classifier to estimate complex nonlinear patterns using HGMD pathogenic and 1000 Genomes Project control variant training set data.	Shihab et al. [57]
deltaSVM	deltaSVM uses a gapped k-mer support vector machine to estimate the effect of a variant in a cell-type-specific manner.	Lee et al. [28]
DeepSEA	DeepSEA uses a multilayered hierarchical structured deep learning-based sequence model to predict functional SNPs with single nucleotide sensitivity using ENCODE and Roadmap Epigenomics data.	Zhou and Troyanskaya [58]

Table 2.1: Online resources for accessing non-coding SNP annotation tools
Abbreviation: HGMD, Human Gene Mutation Database.

Non-coding SNP Validation

RegulomeDB, CADD, FATHMM-MKL, and DeepSEA were all accessed through

their online portals, and run using their default parameters. Variants were submitted in variant call format (VCF) and VCF-like formats. For RegulomeDB, all variants returning a score of 'No Data' were given a score of 7 for downstream analysis. For DeepSEA, functional significance scores were used.

2.9 Notes & Acknowledgments

This chapter was previously published in *Trends in Genetics* (Volume 7, No 1) in January, 2017 [23].

CHAPTER III

Predicting the Effects of SNPs on Transcription Factor Binding Affinity

3.1 Abstract

Genome-wide association studies have revealed that 88% of disease-associated single-nucleotide polymorphisms (SNPs) reside in non-coding regions. However, non-coding SNPs remain understudied, partly because they are challenging to prioritize for experimental validation. To address this deficiency, we developed the SNP effect matrix pipeline (SEMpl). SEMpl estimates transcription factor-binding affinity by observing differences in chromatin immunoprecipitation followed by deep sequencing signal intensity for SNPs within functional transcription factor-binding sites (TFBSs) genome-wide. By cataloging the effects of every possible mutation within the TFBS motif, SEMpl can predict the consequences of SNPs to transcription factor binding. This knowledge can be used to identify potential disease-causing regulatory loci. Availability and implementation: SEMpl is available from https://github.com/Boyle-Lab/SEM_CPP.

3.2 Introduction

To date, genome-wide association studies (GWAS) have identified over 100,000 loci associated with over 200 human diseases and phenotypic traits [69, 36]. Though 95% of known single-nucleotide polymorphisms (SNPs) and 88% of GWAS SNPs fall into non-coding regions of the genome, most genetics studies focus on mutations within coding regions [3, 82]. This large disparity in knowledge gained from big data initiatives is likely due to the more direct interpretability of genic variation even though non-coding variation is also strongly linked to human disease [83, 79]. Identifying non-coding mutations leading to gene misregulation is critical to fully understand GWAS results and their impact on complex and polygenic disorders.

As non-coding GWAS variants are overwhelmingly abundant compared to coding variants, many methods have been developed to prioritize potentially disease-associated mutations in non-coding regions for further study [23]. Generally, these tools focus on known regulatory regions of the genome, relying on variant overlap with experimental annotations, such as regions of open chromatin and transcription factor binding [24, 27, 49]. To date, these computational prioritization tools have assisted in identifying a handful of causal disease mutations from GWAS [62, 61]. However, these tools have only shown up to a 50% concordance rate between predictions, highlighting the need for additional prioritization metrics [23]. One way to improve these predictions is to investigate additional regulatory features to better understand a variant's mechanism of action.

Transcription factor binding sites (TFBSs) are a regulatory feature of particular interest as they make up 31% of GWAS SNPs, yet only comprise 8% of the genome [1]. Mutations in TFBSs influence transcription factor-binding affinity, alter gene

expression, and have been associated with multiple human diseases including cancer and type 2 diabetes, as well as with increased total cholesterol [84, 8, 11, 35, 85, 86]. However, altering different bases within a TFBS have been found to confer different effects on transcription factor binding [2, 41]. This finding has been reflected in cases of human disease, where certain bases in a sequence motif are more correlated with an associated disease than others [87]. Currently, the effect of mutations in a TFBS is estimated using a position weight matrix (PWM), which denotes a transcription factor's binding motif using *in silico* analyses to determine its predominant binding sequence using a competitive binding assay (Figure 3.1A) [88]. PWMs predict where a transcription factor may bind in the genome by acting as its most frequent binding sequence; however they may not recapitulate known binding activity and are not sufficient to predict which mutations within a motif may alter binding affinity [89]. Additionally, using PWMs to predict how a SNP may affect transcription factor binding can be challenging, as PWMs do not contain information on the potential direction of effect of a mutation.

While multiple tools have been developed to predict which mutations may lead to changes in binding affinity, many of these methods rely solely on information from PWMs and are thus subject to similar limitations [90, 91, 92, 93, 94, 95, 96]. More recent methods have incorporated additional measures of binding affinity, including protein-binding microarray data, systematic evolution of ligands by exponential enrichment (SELEX) data and/or chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) data [97, 98, 99, 28, 100, 101, 26]. These methods represent a marked improvement over a strict input of PWMs, however they have their own pitfalls. As protein-binding microarray and SELEX data are generated outside of a native cell context they may not represent patterns of true intercellular binding.

In addition, the majority of these models output de novo motifs similar in style to PWMs, which are not informative to direction of effect of a mutation. One tool of particular interest, the Intragenomic Replicates (IGR) method, was developed as a way to investigate FOXA1 involvement in breast cancer using GWAS data [102]. This method compares TFBSs containing putatively deleterious mutations to their wild-type counterparts using genome-wide ChIP-seq data to estimate predicted changes to transcription factor-binding affinity. The predictions generated by IGR were found to be highly correlated with ChIP-qPCR results and were successfully used to identify a risk allele associated with a 5-fold change in gene expression in breast cancer. IGR represents a marked improvement over other methods due to its specific calibration of variants to ChIP-seq data, an endogenous source of transcription factor-binding affinity information. Currently, IGR exists only as a method designed to probe individual mutations and must be reconstructed for each new mutation and transcription factor. However, the premise of using ChIP-seq data to predict transcription factor binding could be expanded to more quickly and accurately predict TBFS mutations.

In order to improve current methods to be applicable to a wide range of transcription factors and to better predict which mutations within TFBSs may lead to changes in binding affinity, we have developed a new method: the SNP effect matrix pipeline (SEMpl). Our method uses endogenous ChIP-seq data and existing variants genome-wide similar to the IGR method, however SEMpl also includes a catalog of kmers separated by a single base change from a TFBS motif, allowing it to provide an estimate of the consequence of every possible mutation in a TFBS. We call these as SNP effect matrices (SEMs, Figure 3.1). Here, we demonstrate that SEMs recapitulate known motifs, are robust to input data and cell type, and are better at predicting changes to transcription factor-binding affinity than the current

standard, PWMs. By developing SEM scores, we aim to improve the prioritization of non-coding GWAS variants for further experimental validation, expand the understanding of non-coding genomic variation and further technology toward developing tools for personalized medicine.

3.3 Materials and methods

Usage/accessibility

SEMpl is open access and can be downloaded from github: https:// github.com/Boyle-Lab/SEM_CPP.

SNP effect matrix pipeline

SEMpl utilizes three types of experimental evidence to make its predictions: ChIP-seq data, which provides a transcription factor's endogenous binding in the genome; DNase I hypersensitive site (DNase-seq) data, which represents regions of open chromatin where transcription factors are known to function and PWMs, which denote previous knowledge of the binding pattern of transcription factors (Figure 3.1). We obtained ChIP-seq and DNase-seq data from the ENCODE project and PWMs from the JASPAR, Transfac, UniPROBE and Jolma databases [1, 103, 104, 92, 52].

SEMpl first enumerates a PWM of interest into a list of kmers using a permissive cutoff P-value threshold of 4 5 using the software transcription factor matrix P-value (TFM-PVALUE) (Figure 3.2A) [105]). This first list of kmers, referred to as the endogenous kmer list, represents sequences where the transcription factor of interest has an increased likelihood of binding. To observe additional sequences which may show distinct binding preferences, SEMpl next takes the endogenous kmer list and

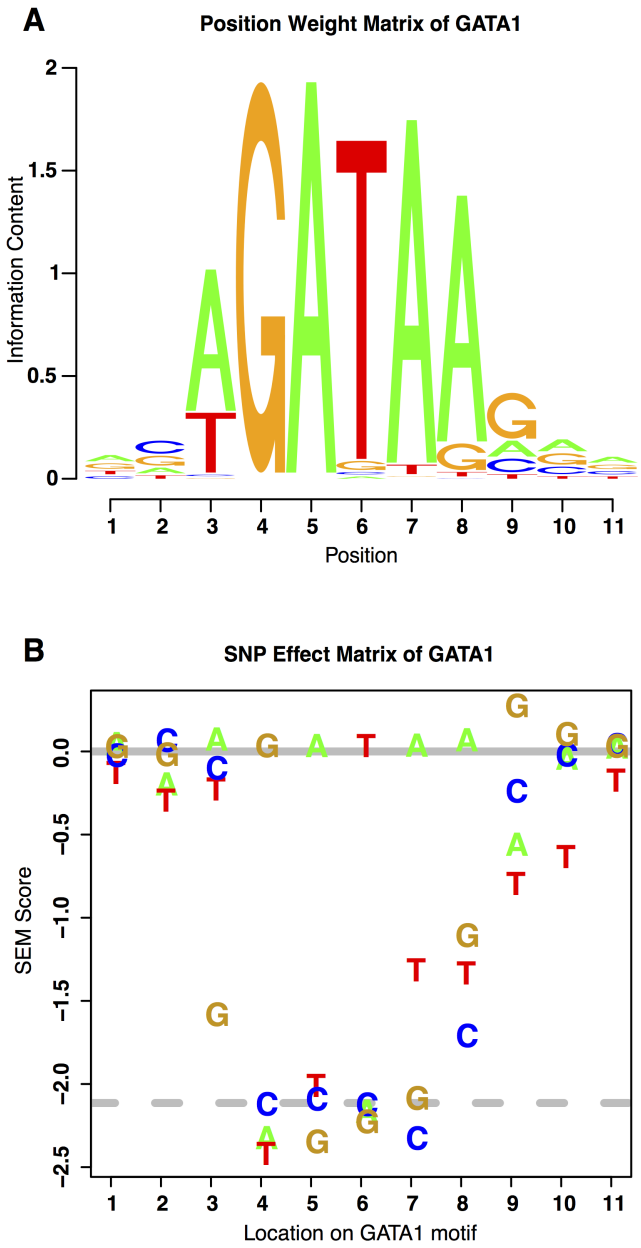


Figure 3.1: PWM versus SEM of transcription factor GATA1. (A) The PWM can be read as likely nucleotides along a transcriptions factor's motif. (B) Similarly, the SEM can be read as nucleotides along a motif, but with additional information about the effect any given SNP may have on transcription factor-binding affinity. The solid gray line represents endogenous binding, the dashed gray line represents a scrambled background. We define anything above the solid gray line as predicted to increase binding on average, anything between the two lines as decreasing average binding and anything falling below the dashed gray line as ablating binding on average.

simulates all possible SNPs in silico to create lists of mutated kmers (Figure 3.2B). For example, by changing all bases in position 6 to a G nucleotide in every kmer in the endogenous kmer list, SEMpl creates a mutated kmer list for G in position 6. These lists of mutated kmers are then aligned to the hg19 reference human genome in regions of open chromatin using bowtie, as determined by DNase-seq [106]. The ChIP-seq score is then calculated as the highest signal value over the region 50bp before and after the aligned site (Figure 3.2C). Next the SEM score for each position is computed as the log2 of the average ChIP-seq signal to endogenous signal ratio for the mapped kmers for each mutated kmer list. Taken together, the SEM scores for each base form a matrix for each nucleotide at every position along the motif. Scores can be evaluated at individual nucleotides, or calculated across a full-length kmer by adding the nucleotide score for each position along the motif, similar to a PWM.

The above process is repeated, using a slightly more stringent TFM-PVALUE cutoff of $4^{-5.5}$ to generate kmers, until convergence using an estimation maximization (EM)-like method in order to correct for differences arising from unique starting kmers (Figure 3.2D, Figure 3.3). This process continues until the number of kmers from the endogenous kmer list does not change or until 250 iterations, with the average run converging by iteration 117. To control for poor quality data and to identify background levels of binding, a final kmer list of randomly scrambled endogenous kmers is included to represent a random baseline where transcription factor binding would not be expected to occur (displayed as a dashed gray line on an SEM plot). Finally, we define scores above 0 as predicted to increase binding on average, scores between 0 and the scrambled background as decreasing average binding on average and scores falling below the scrambled background as ablating binding on average.

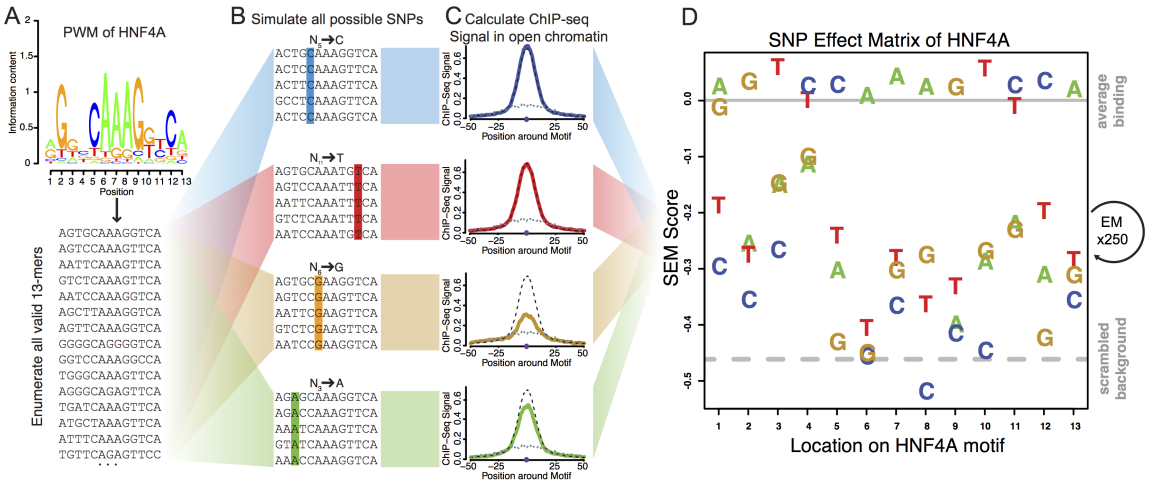


Figure 3.2: SEM methods pipeline. (A) All kmers with a PWM score below the TFM-PVALUE are generated for a single transcription factor. (B) All possible SNPs are introduced in silico for each kmer. (C) All enumerated kmers are then aligned to the genome, and filtered for regions of open chromatin by DNase-seq. The average ChIP-seq scores are then calculated for each alignment (dashed line represents endogenous binding, dotted line represents scrambled background). (D) Final SEM scores are log2 transformed and normalized to the average binding score of the original kmers (solid gray line). A scrambled baseline, representing the binding score of randomly scrambled kmers of the same length is also added (dashed gray line). Once a SEM score is calculated, the output can be used to generate a new PWM. This iterative process can correct for disparities introduced by the use of different starting PWMs. The HepG2 cell line data were used for the ChIP-seq and DNase data for HNF4a.

SEMpl output files include error messages during the run (.err), the cache, a tally of kmer similarity between iterations (kmer_similarity.out) and an output file containing information on run time and where the program is in the run (.out). Additionally, within each iteration, output files include the alignments for the SNP kmer lists (alignment folder) and endogenous and scrambled kmer lists (baseline folder) which include the aligned loci and ChIP-seq signal. A quality control file is also provided within each iteration file that provides the number of kmers mapped within the iteration, as well as a –log10(P-value) representing the average of 100 t-tests from 1000 randomly chosen kmers from the SNP signal files versus 1000 randomly chosen kmers from the scrambled signal file. We used a threshold of 2.5 to report confidence in a SEM run. Resulting aligned loci and ChIP-seq values are stored in

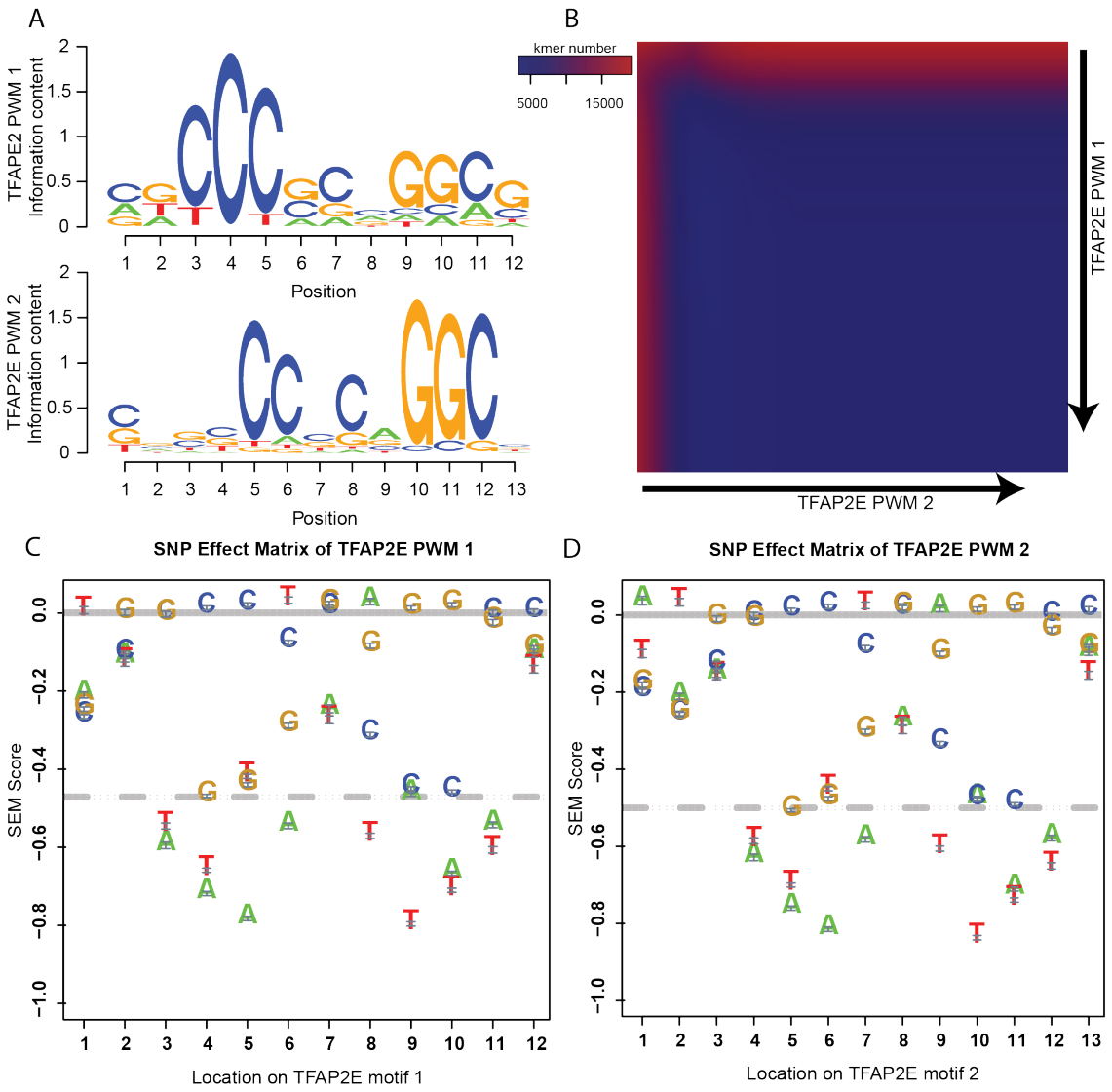


Figure 3.3: Different starting PWMs yield highly similar SEMs. A. Two starting PWMs were chosen, representing distinct binding profiles and motif length (TFAP2E PWM 1, M00189; TFAP2E PWM 2, M00915). B. Both PWMs were run using SEMpl, and similarities between kmers generated from PWMs were tracked over 250 iterations. Red represents more differences in number of kmers between the two sets and deep blue represents fewer differences in number of kmers between the two sets. C&D. Final SEMs from starting with PWM 1 or PWM 2 appear highly similar, with the SEM of PWM 2 containing an extra nucleotide in position 1 of the motif.

a cache, which allows for a quick lookup of nonunique kmers without realignment. SEMpl options include –readcache, which can be used to speed up a run for which a cache has already been created. SEMpl is written in C++ and R. PWMs were created using the R package seqLogo [107].

Scoring a variant or sequence with a SEM

Scoring variants or sequences using SEMpl are as straightforward as scoring using a PWM. A score can be computed in two ways. First, a single base change can be scored by subtracting the wild-type nucleotide score from the variant score using the SEM matrix to determine the total predicted difference between the two nucleotides. Second, a kmer sequence can be scored in a manner very similar to a PWM. Because the matrix is log transformed, the score of each nucleotide can be added to reflect the predicted binding of the full sequence. In this way the effect of multiple variants can be calculated for a single sequence. In either case, the final value represents the expected change compared to endogenous binding levels.

Correlation with ChIP-seq data

All possible kmers from the original transcription factor PWMs were generated. For each unique kmer, average ChIP-seq signal and standard error were calculated. PWM, SEM, DeepBind and LS-GKM scores were calculated for each kmer. DeepBind scores were calculated from precomputed models, and LS-GKM scores were computed using the options l=10 and k=6 for motifs with length $\geq$10—as recommended by the author. For LS-GKM motifs length 9, I=9 and k=6, and motifs length 8 were run using l=8 and k=5. Correlations cutoffs were calculated for PWMs above the standard TFM-PVALUE cutoff (P-value=4^{-8}) typically used for PWM visualization. Correlation cutoffs for SEM, DeepBind and LS-GKM scores were defined as the average scrambled baseline across all iterations for a single transcription factor run.

SEM correlation across runs

SEM outputs from different starting ChIP-seq or PWM data were compared using least square regression in R. Overlapping DNase-seq peaks were downloaded from ENCODE and calculated using bedtools [108]. SEMpl runs from the same cell type, and therefore using the same DNase dataset, share 100% DNase peak overlap.

Allele-specific CTCF-binding pattern analysis

Allele-specific binding sites were defined as loci containing one or more heterozygous SNPs while showing significant differences in ChIP-seq signal from two alleles. We applied the AlleleDB pipeline to count the number of ChIP-seq reads from two alleles respectively for each heterozygous site and identified 468 allele-specific binding sites at an FDR of 5% [109]. CTCF ChIP-seq data from GM12878 cell line was used in this analysis (accession number: ENCSR000DZN). For all heterozygous sites within CTCF ChIP-seq peaks in GM12878 cell line, 240 of them also have matching CTCF PWMs, which we further used for the comparison of SEM and PWM scores. For those 240 heterozygous sites, we calculated the allelic ratio defined by the ratio between the number of ChIP-seq reads from the maternal allele and the total number of reads from two alleles. We then evaluated the correlation between the change of SEM or PWM scores and allelic ratios.

Electrophoretic mobility shift assay (EMSA) analysis

The DNA-binding domains of CTCF (F1–F9) were amplified from Addgene plasmid 102859 and cloned into a bacterial expression vector with a GST tag (pGEX4T) [110]. This construct was transformed into BL21(DE3) cells. 1L LB liquid bacteria cultures were induced by 0.25mM IPTG at OD600=0.6 and incubated at 12 C for

24h. Cells were lysed by sonication, and GST-CTCF was pulled down by a glutathione column. Following five washes with wash buffer (20mM HEPES-KOH, pH 7.2, 150mM KCl, 0.05% NP-40, 10% glycerol), the sample was cleaved by thrombin and run through a column, resulting in purified, cleaved CTCF (F1–F9) protein (Figure 3.4A).

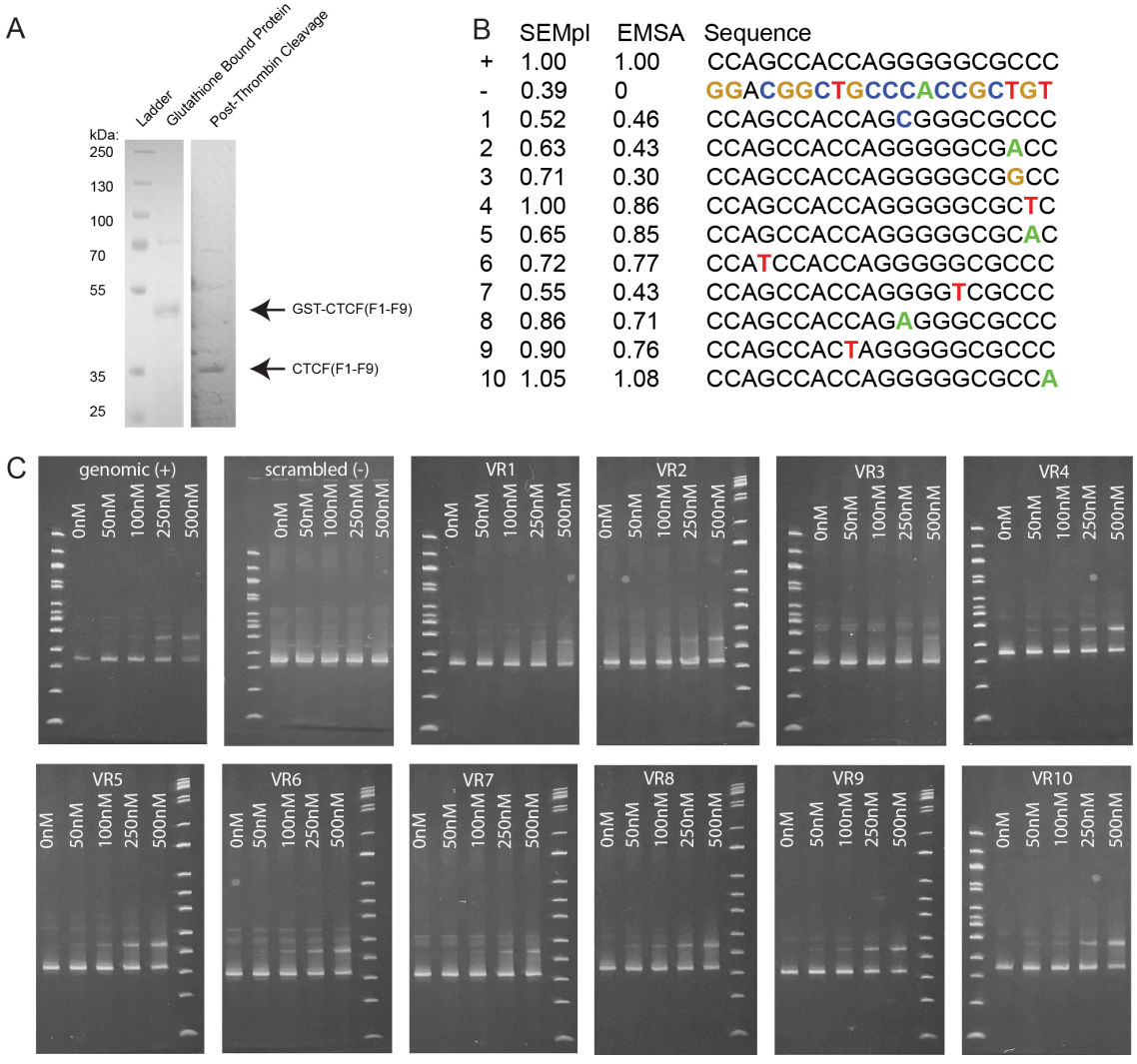


Figure 3.4: Electrophoretic mobility shift assay (EMSA) for CTCF. A. Coomassie-stained gel showing GST-tagged CTCF protein DNA binding domain (F1-F9) fragment and thrombin-cleaved CTCF protein fragment. B. Table of variable region sequences alongside SEMpl predictions and EMSA scores. All scores EMSA scores normalized to the genomic (+) sequence and scaled between 0 and 1. Color corresponds to the SNP made in the variable region. C. Gel Images of variable region EMSAs. 50nM variable regions were incubated with 0nM, 50nM, 100nM, 250nM, and 500nM CTCF protein fragment. All ladders on the left side of their gel are 100bp (NEB, N3231S) and all ladders on the right side of their gel are 1kb plus (Invitrogen, 10787018).

For our EMSA analysis, we tested a 20-bp genomic binding region to CTCF flanked by 200 bp upstream and downstream of endogenous sequence (hg19, chr9: 135045357–135045377). We introduced mutations to create 10 variable regions containing a single mutation and one scrambled region. We completed EMSAs as previously reported [111], incubating 50 nM DNA fragments with 0, 50, 100, 250 and 500 nM purified CTCF protein fragments for 30min. EMSA reactions were then run on 4–12% TBE gels (EC62352BOX) for 3h at 80 V and 4 C. EMSA analysis was completed as previously reported using densiometric scanning by ImageJ and an Excel Solver Package [112, 113]. EMSA scores were normalized to the genomic background (+) and scaled between 0 and 1.

3.4 Results

SEM scores better recapitulate endogenous binding than PWMs

SEM scores are expected to be more representative of endogenous binding patterns than PWMs as these predictions are generated using an endogenous measure of genome-wide binding affinity. We demonstrate this by correlating SEM and PWM scores across full-length kmers for transcription factor FOXA1 to their average ChIP-seq signals at corresponding sequences genome wide (Figure 3.5). When comparing predictions with experimentally generated binding affinity data above standard cutoffs, SEMs had a stronger correlation than PWMs (SEM: R^2=0.66, PWM: R^2=0.24), demonstrating our predictions represent a more robust measure of endogenous binding affinity. This pattern holds true when allowing a very lenient PWM cutoff of 11 (R^2=0.28) as well as for the entire datasets (SEM: R^2=0.19; PWM: R^2=0.03) (Figure 3.6).

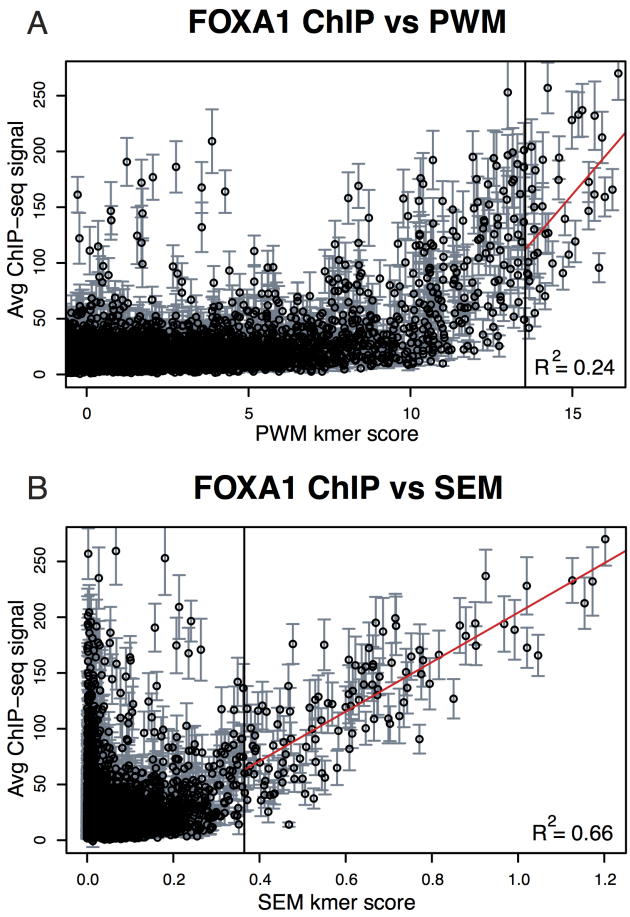


Figure 3.5: SEMs show a better correlation with whole kmer ChIP-seq signal (B, R^2=0.66) than PWMs (A, R^2=0.24). The line dividing the plot represents a standard cutoff for PWM visualization (P-value=4^{-8}). Coefficient of determinations (R^2) were calculated to the right of the vertical lines, representing the TFM-PVALUE cut-off for PWMs and the average scrambled background cutoff for SEMs (0.36 for FOXA1). SEM values are displayed as 2n for visualization purposes. PWM values only shown >0, a full plot can be found in Figure 3.6

These findings indicate that SEM plots better recapitulate known patterns of transcription factor binding beyond the information detailed in a PWM. Of note, there are cases where the PWM shows approximately equal information content for distinct bases sharing a position, yet the SEM plot reveals a wide margin of binding differences between the two bases fueled by differences in predicted direction of effect on binding affinity (i.e. position 3 or 10 of HNF4a in Figure 3.2).

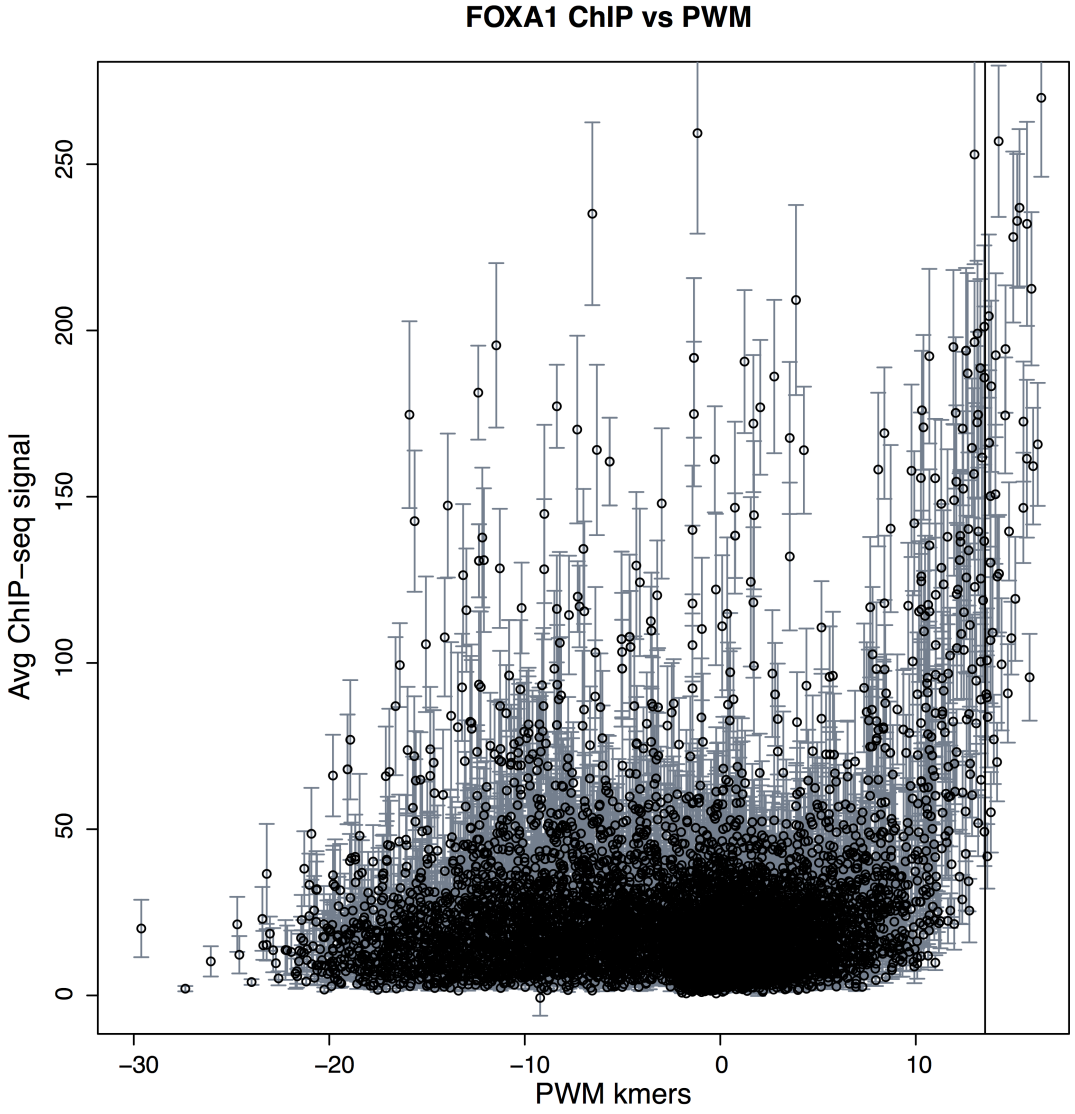


Figure 3.6: Correlation of PWM scores for full kmers versus average ChIP-seq signal. All possible kmers from a starting FOXA1 PWM were included. This is an extension of Figure 3.5, with x-axis bounds from -30 to 15.

Ubiquitous transcription factors show cell type and dataset independence

To determine if SEM results show a dataset-specific dependence, we evaluated the transcription factor FOXA1 using ChIP-seq data from two different ENCODE datasets gathered in the same HepG2 cell line (ENCFF658RGX; ENCFF898FCL) (Figure 3.5). We found nearly identical SEMpl outputs (P-value=4.14e-56, RMSD=0.0178)

using least-squares regression analysis.

We next expanded this to investigate if SEM results were dependent on the cell line used and thus included three additional ChIP-seq datasets (ENCFF699KBP; ENCFF845PAS; ENCFF723DLM) from distinct cell types (Figure 3.7). It is important to note that while some of the regions tested in the cell lines are at the same locations, there are large differences in the open chromatin regions (and thus site accessibility) across these cell types, often with >50% unique sites between cell types (bottom half of Figure 3.7). We saw high levels of correlation using these additional cell types, with R^2 values over 0.97 for HepG2, A549 and T47D (P-values <1e-32, RMSD <0.0717). We also saw this trend between SEMs run on different cell lines for additional transcription factors including in MYC, NKFB1 and FOS, suggesting that for ubiquitous transcription factors, we expect there to be no appreciable difference between SEMpl outputs (Figures 3.8-3.10).

It has been proposed that there may be binding affinity differences between cell types when a transcription factor has known cell type-specific functions or cofactors. To address this, we investigated the protooncogene MYC, which encodes for the transcription factor c-myc known to have distinct functions and cofactors between differing cell types [114]. Interestingly, we found that c-myc yielded a highly similar pattern between almost all cell types observed, but a distinct SEM plot in HeLa cells that cannot be explained by low data quality (Figure 3.8). This suggests that SEMpl can also be used to identify transcription factors that have distinct cell type-specific functions. However, this seems to be the exception rather than the rule as the majority of SEMs we observed were cell-type agnostic.

Finally, we asked if the starting PWM for a TF would influence the final SEM output. We found no appreciable difference in SEMpl outputs when using different

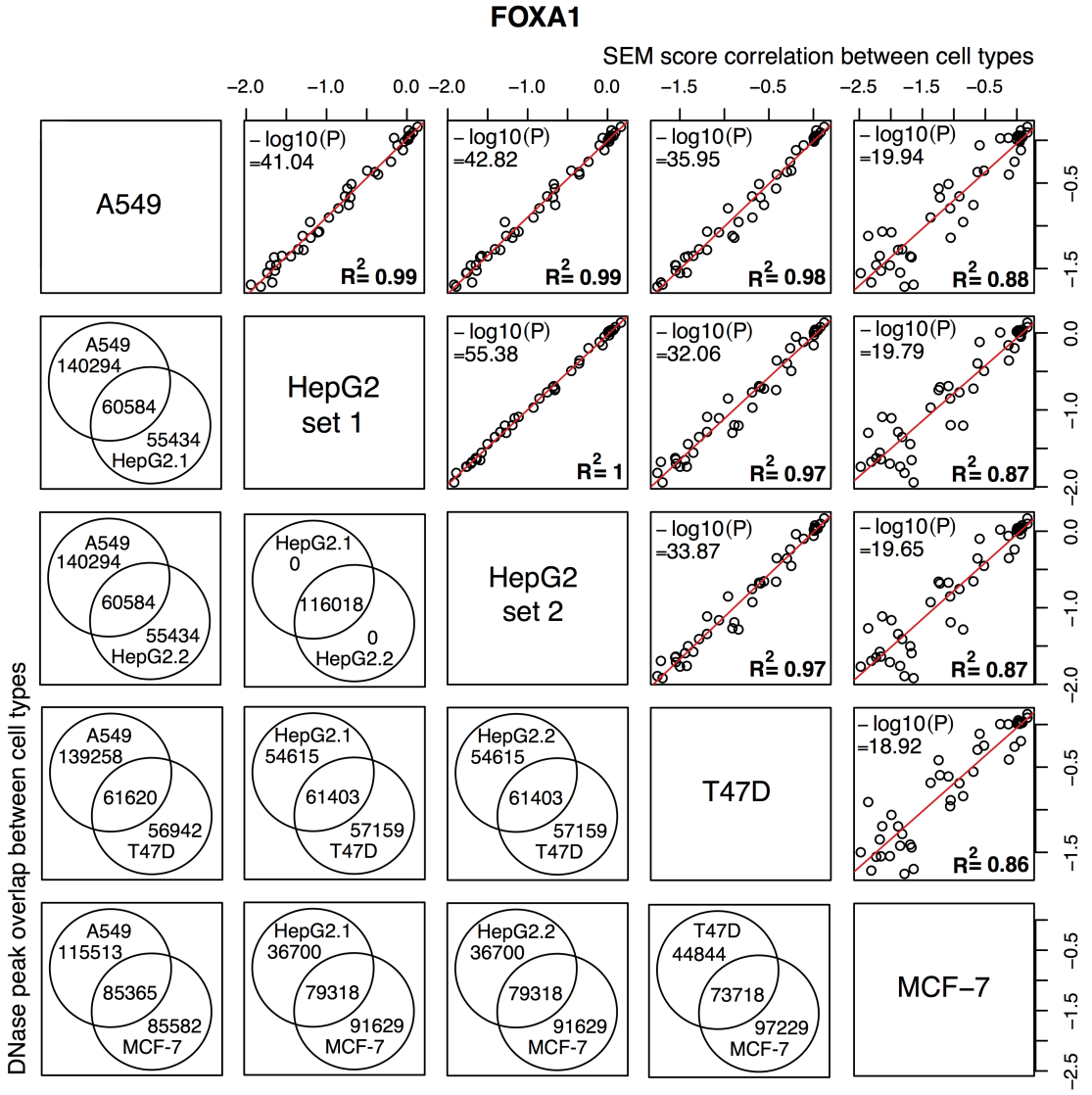


Figure 3.7: Different ChIP-seq input produce similar SEMs. The top right half of the table shows a least square regression analysis which reveals that FOXA1 SEMs are highly correlated across four cell types and one pair of biological replicates with correlations between samples ranging from $R^2 = 0.86$ and $R^2 = 1$. The bottom left half of the table shows overlapping DNase peaks between cell types. A549, lung carcinoma cell line HepG2, hepatocellular carcinoma cell line T47D, breast tumor cell line MCF-7, breast adenocarcinoma cell line.

starting PWMs, given that the starting PWMs represent the general binding of the transcription factor of interest (Figure 3.3). However, certain PWMs and/or datasets do not contain enough information about the binding of a TF and so do not produce any significant enrichments in the final SEM output and are thus discarded.

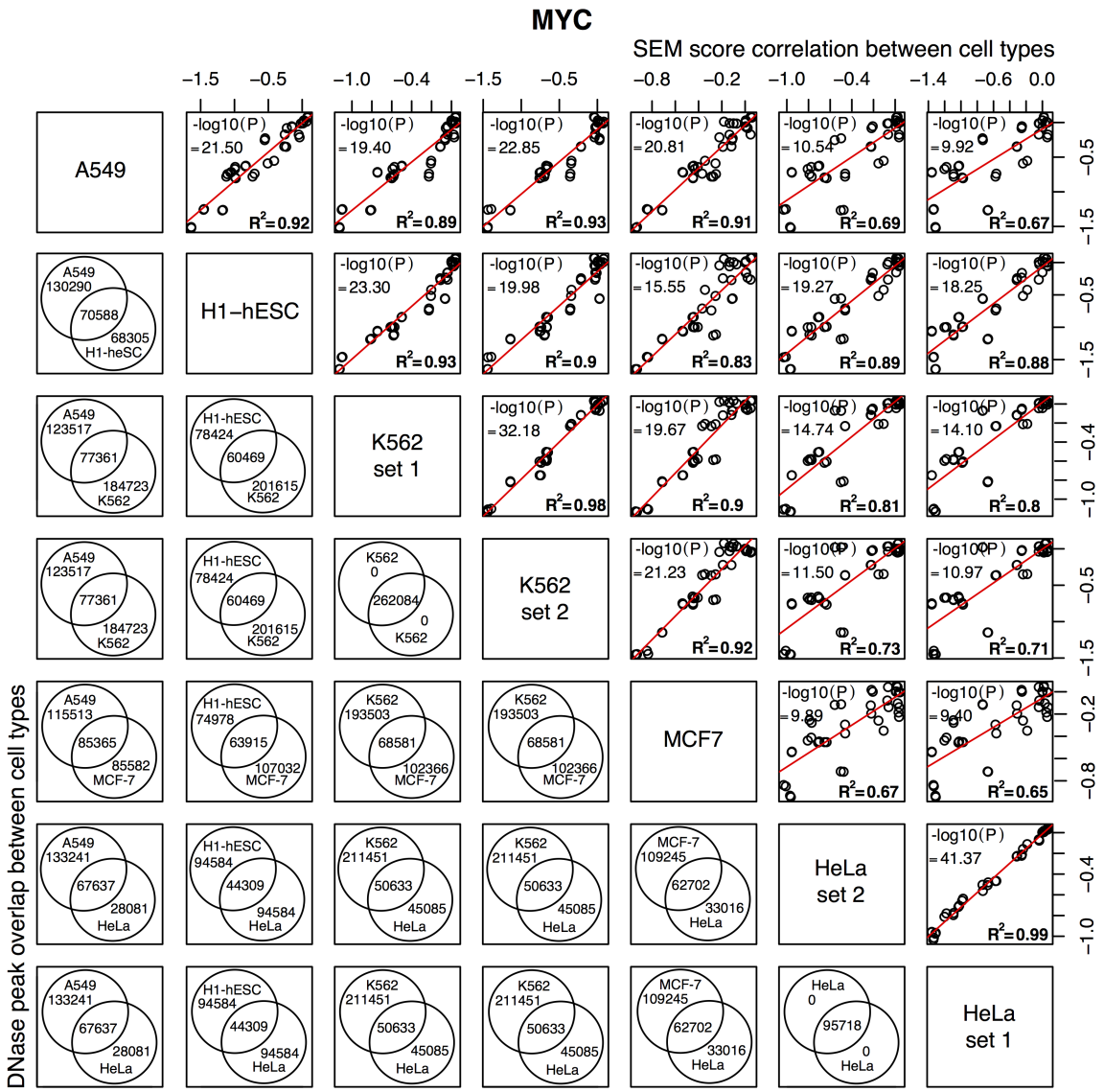


Figure 3.8: MYC shows cell-line specific binding affinity across different ChIP-seq input data (HeLa v. others). The top right half of the table shows a least square regression analysis, while r-squared analysis and p-values can be found on the bottom left.

SEMpl recapitulates known allele-specific binding patterns

Allele-specific binding differences in non-coding regions of the genome have long been associated with regulatory sequence [2, 41]. To compare SEM scores against known allele-specific binding data, we annotated heterozygous sites in the GM12878 cell line with ChIP-seq read counts from two alleles using ENCODE CTCF ChIP-seq

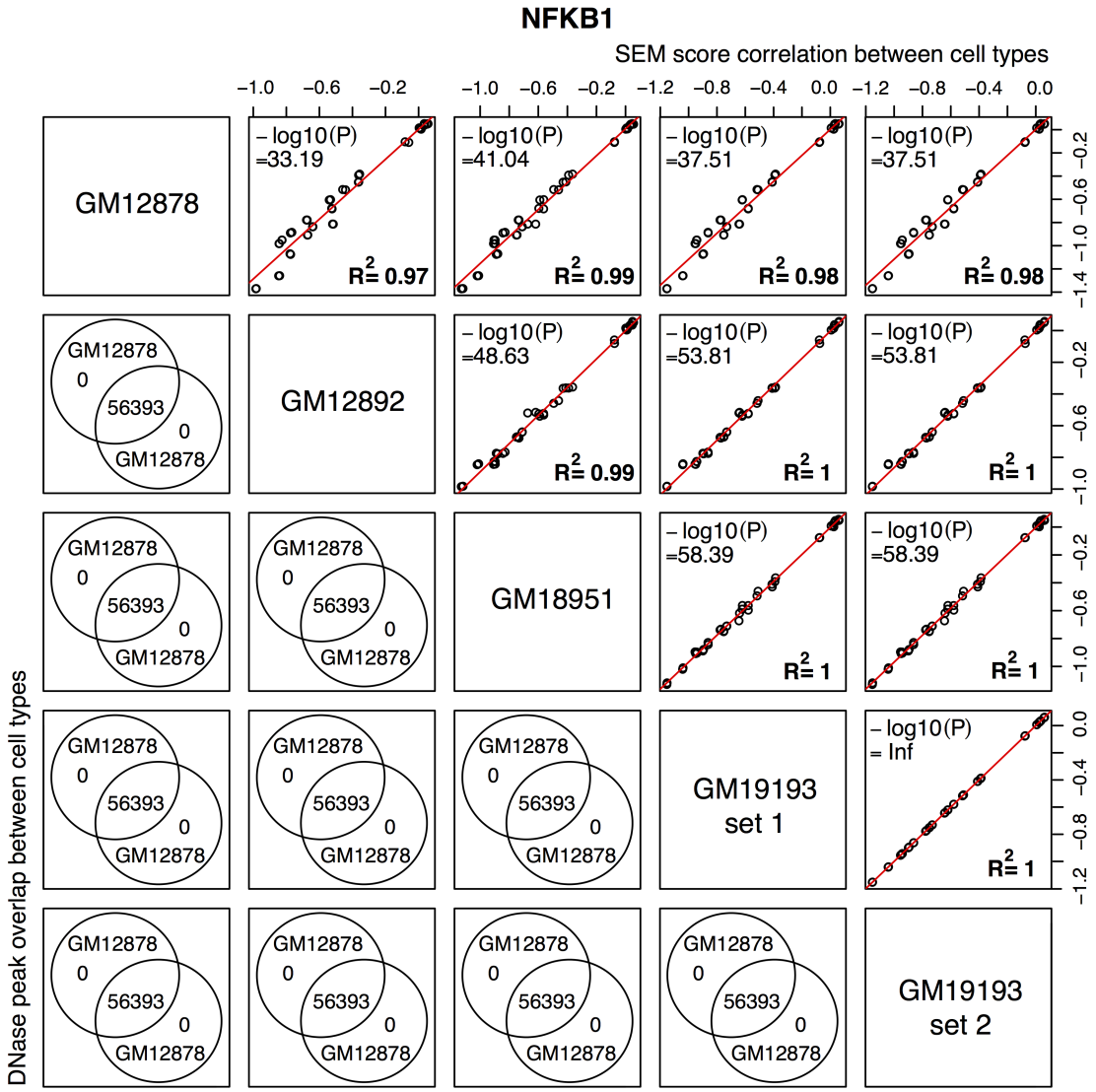


Figure 3.9: NFKb shows robust correlations across different ChIP-seq input data. The top right half of the table shows a least square regression analysis, while r-squared analysis and p-values can be found on the bottom left.

datasets. Least-squares regression analysis of SEM or PWM score changes against ChIP-seq signal changes of these 240 heterozygous sites in CTCF-binding sites revealed a higher correlation for SEM score changes with an R^2 of 0.50 compared to a PWM R^2 of 0.41 (Figure 3.11). We also observed a more dispersed distribution of SEM score changes, where the allele-specific binding sites have overall larger changes between two alleles (red points in Figure 3.11). These indicate that the SEM score

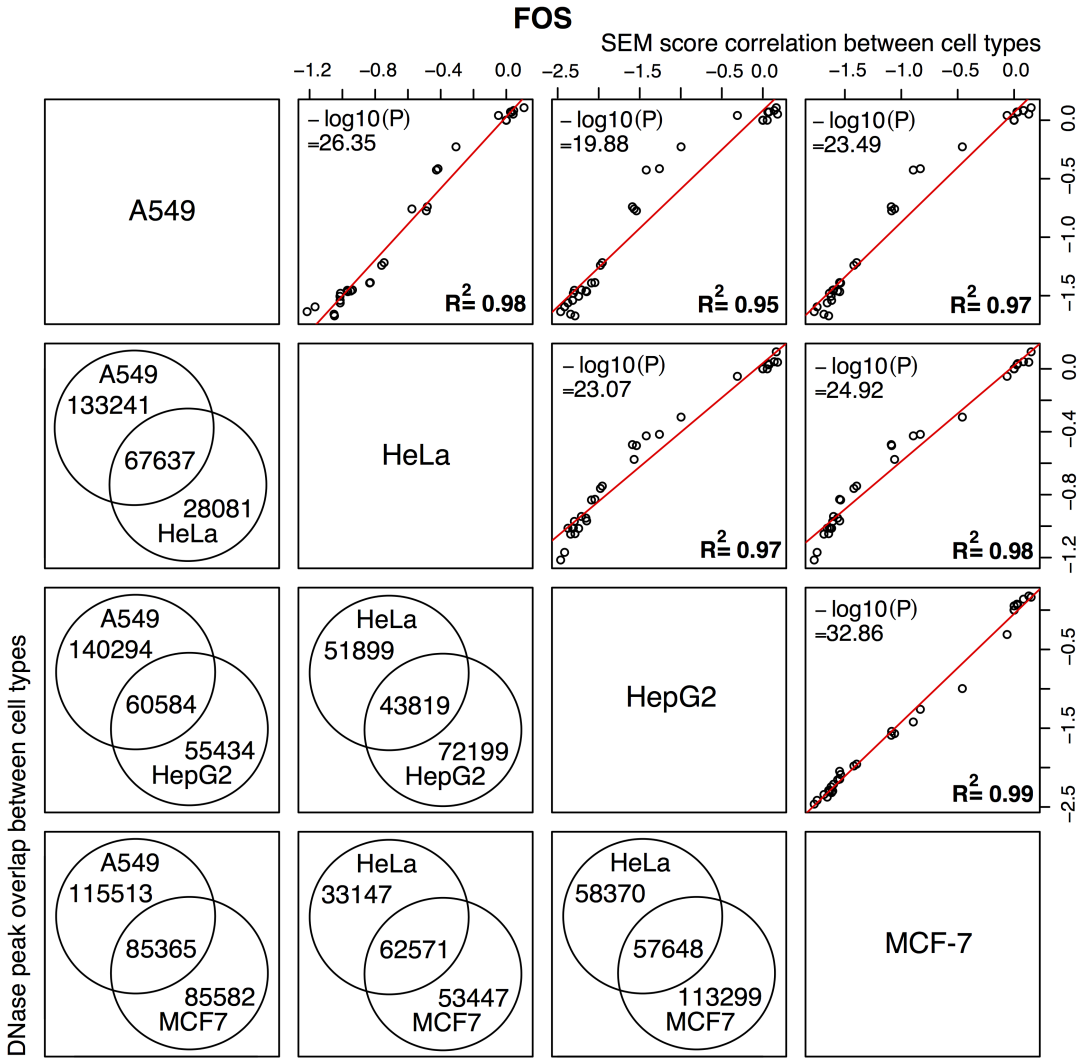


Figure 3.10: FOS shows robust correlations across different ChIP-seq input data. The top right half of the table shows a least square regression analysis, while r-squared analysis and p-values can be found on the bottom left.

is more able to capture the change of TF-binding affinity compared to PWM.

To validate that SEMpl scores accurately predict transcription factor-binding affinity changes *in vitro*, we compared SEMpl scores to previously generated ChIP-qPCR data, which measures endogenous transcription factor-binding affinity [102]. ChIP-qPCR was generated from 10 allele-specific FOXA1-binding sites in the genome. Regression analysis comparing SEMpl scores to changes in transcription factor bind-

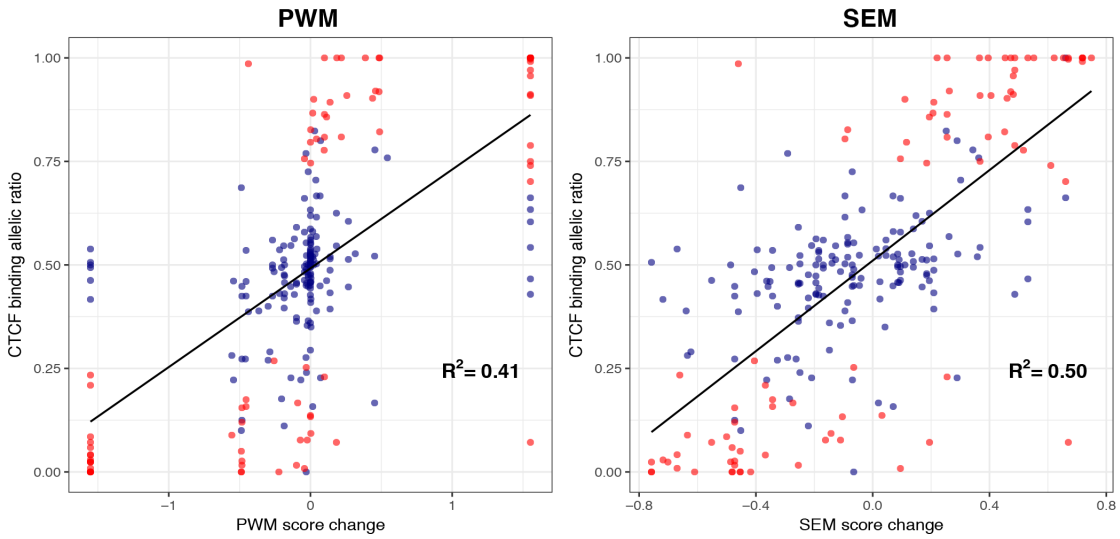


Figure 3.11: SEMs reflect allele-specific CTCF-binding patterns. Linear regression reveals a higher correlation between SEM score change and binding affinity change in two alleles of heterozygous sites (R^2=0.50) than PWM scores (R^2=0.41). Allele-binding affinity change was measured by allelic ratio, which is the ratio between CTCF ChIP-seq read counts from maternal allele and total read counts from two alleles. Allele-specific binding sites (red/light gray points) generally have larger changes on SEM scores.

ing by ChIP qPCR analysis reveal that SEM scores are a better predictor of SNP changes ($R^2 = 0.64$) than PWMs ($R^2 = 0.44$) (Figure 3.12).

We examined SEMpl predictions further by comparing them to *in vitro* binding data generated by EMSA of purified protein of the DNA-binding domains of CTCF to engineered DNA consensus sequences. EMSAs of 10 CTCF-binding sites containing a mutation, which we define here as variable regions, compared to a known CTCF-binding site along with the endogenous sequence and scrambled background reveals a better correlation with SEM predictions ($R^2 = 0.76$) than PWM predictions ($R^2 = 0.65$) (Figure 3.13A, Figure 3.4). This is further supported by comparing SEM and PWM scores to previously published EMSA data for the mouse transcription factor FoxA1 [111]. This analysis showed a marked improvement of SEM scores ($R^2 = 0.75$) compared to PWM scores ($R^2 = 0.6$), and machine learning models DeepBind ($R^2 = 0.66$) and LS-GKM ($R^2 = 0.67$), and suggests that the SEMs of highly

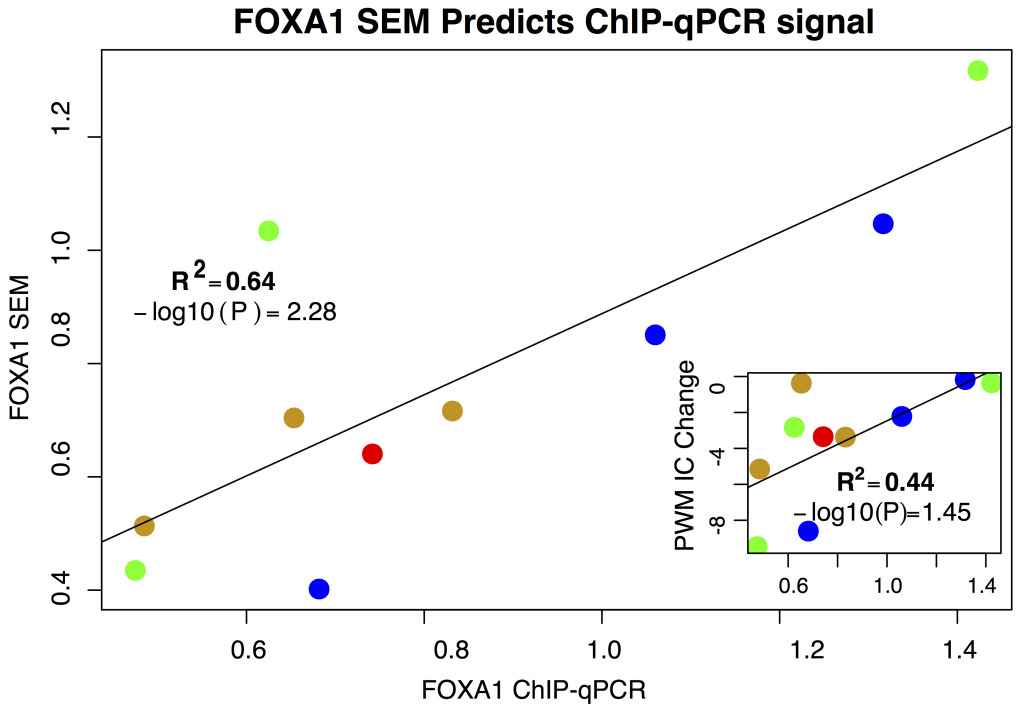


Figure 3.12: SEMs scores are a better predictor of transcription factor binding changes from SNPs than PWMs when compared to ChIP qPCR analysis for FoxA1 [102]. ChIP-qPCR data for 10 FOXA1 SNPs from the IGR paper was used as an endogenous measure of binding affinity change. Colors represent the SNP change (green, A; blue, C; yellow, G; red, T). The SEM was generated using HepG2 cell line data for the ChIP-seq and DNase. PWM change is measured in change to information content (IC).

conserved transcription factors may be comparable between species (Figure 3.13B) [97, 28]. Together, these results suggest that SEMpl has the ability to return biologically meaningful results and can be used to predict the direction and magnitude of allele-specific changes.

SEMpl predictions agree with experimentally validated SNPs from the literature

To verify that SEMpl would allow researchers to identify variants potentially leading to transcription factor-binding changes associated with gene expression changes, we validated our method against four published TFBS SNPs found to disrupt tran-

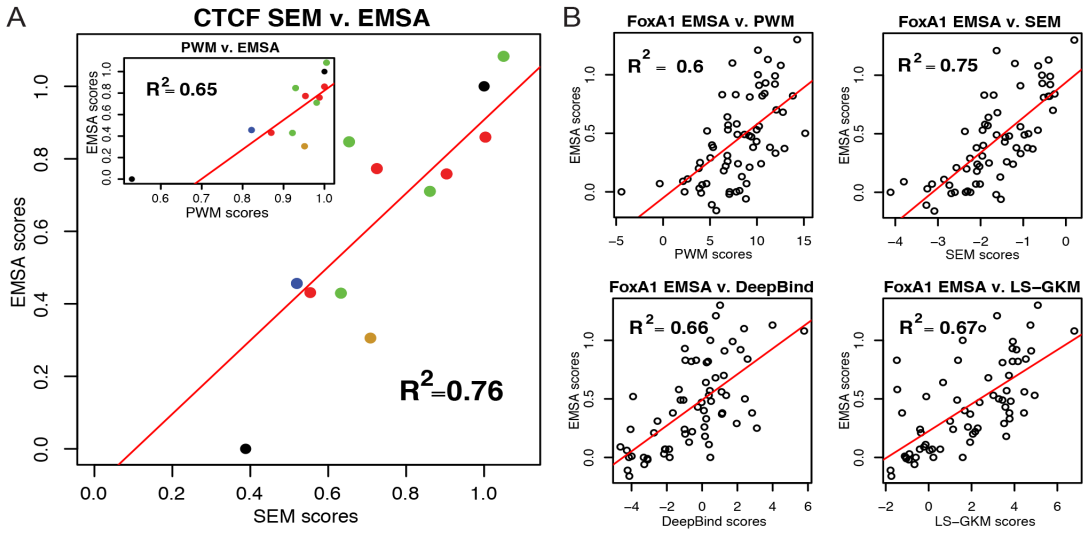


Figure 3.13: SEMpl scores agree with *in vitro* transcription factor-binding results. (A) Electrophoretic mobility shift assay (EMSA) for CTCF correlated to SEMpl and PWM predictions. Correlations are calculated without the inclusion of the genomic and scrambled controls (black points). Additional colors correspond to the SNP change made to the variable region. (B) FoxA1 EMSA data from Levitsky et al. correlated to PWM, SEM, DeepBind and LS-GKM predictions [111].

scription factor binding (Figure 3.14). In most cases, we found that SEMpl predictions agreed with the direction of the validated changes, as well as the magnitude, when available. For example, a T to G change in position 12 of a TCF7L2-binding site was found to increase binding affinity by 1.3-fold by mass spec [35], where SEMpl predicted a 1.27-fold increase. Only one of the four SEMpl predictions that we identified did not match the experimentally determined variant. This C/T allele in position 11 of a FOXA2-binding site was predicted to decrease binding affinity by FAIRE-seq, however SEMpl predicted no difference in binding between the two alleles (data not shown). Interestingly, PWMs also predicted no difference in binding between the two alleles, suggesting additional factors may be at play.

We also compared SEMpl predictions to predicted variant effects measured through a massively parallel reporter assay (MPRA) [32]. We found a correlation between these previously published expression changes and SEM score changes (Figure 3.15).

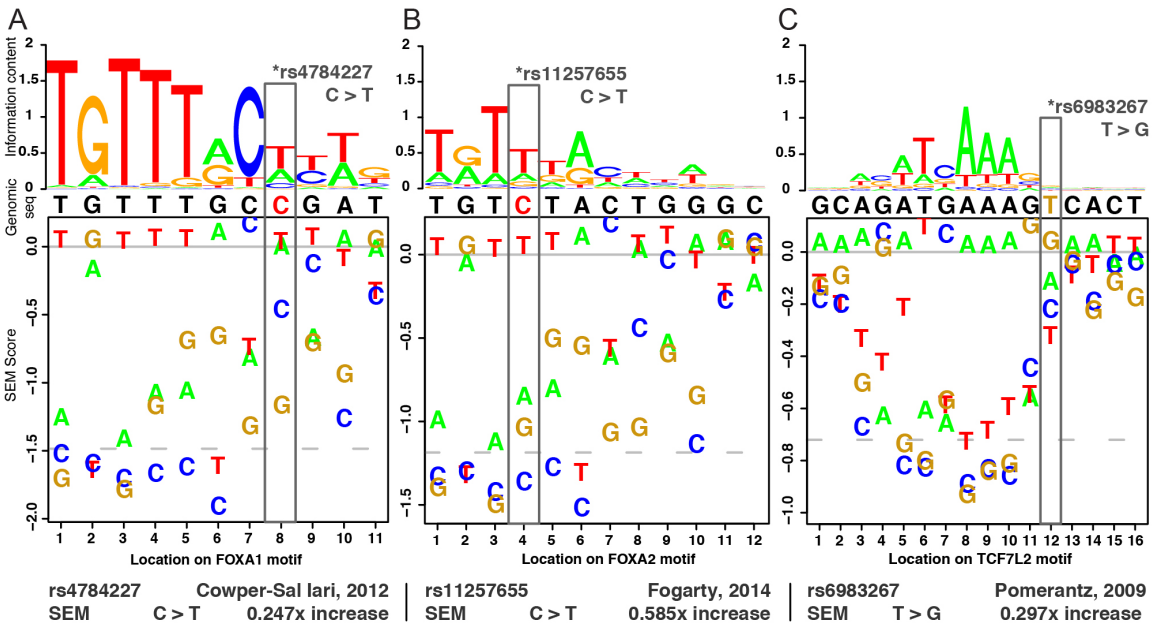


Figure 3.14: Known variants affecting transcription factor binding affinity [102, 84, 35].

However, this relationship was not as strong ($R^2 = 0.23$), though still outperforming PWMs ($R^2 = 0.16$), possibly due to the nonlinear relationship between transcription factor binding, regulatory element use and gene expression.

SEMpl outperforms other methods in predicting changes to transcription factor binding

In order to compare SEMpl to current state-of-the-art methods, we compared SEMpl and PWMs to methods utilizing machine learning able to predict the consequence of variants to transcription factor binding, DeepBind and LS-GKM [97, 28]. Both tools use models trained on ChIP-seq datasets to generate predictions of function variation. Here, we compared scores for all methods (PWM, SEMpl, Deepbind and LS-GKM) against ChIP-seq scores for all kmers from 13 transcription factors (Figure 3.16).

Using a performance comparison, we found that SEMpl better correlates with

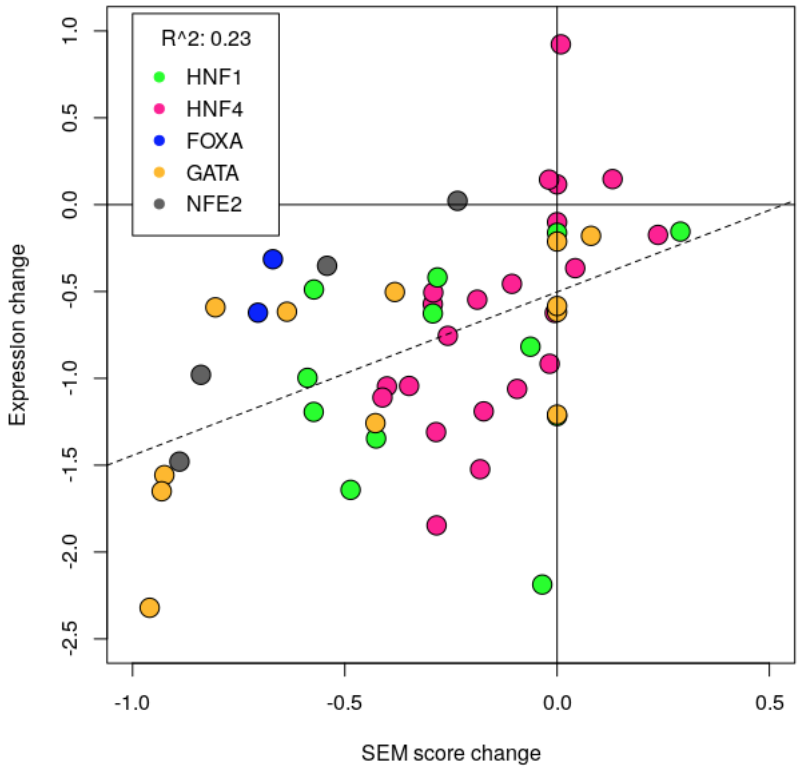
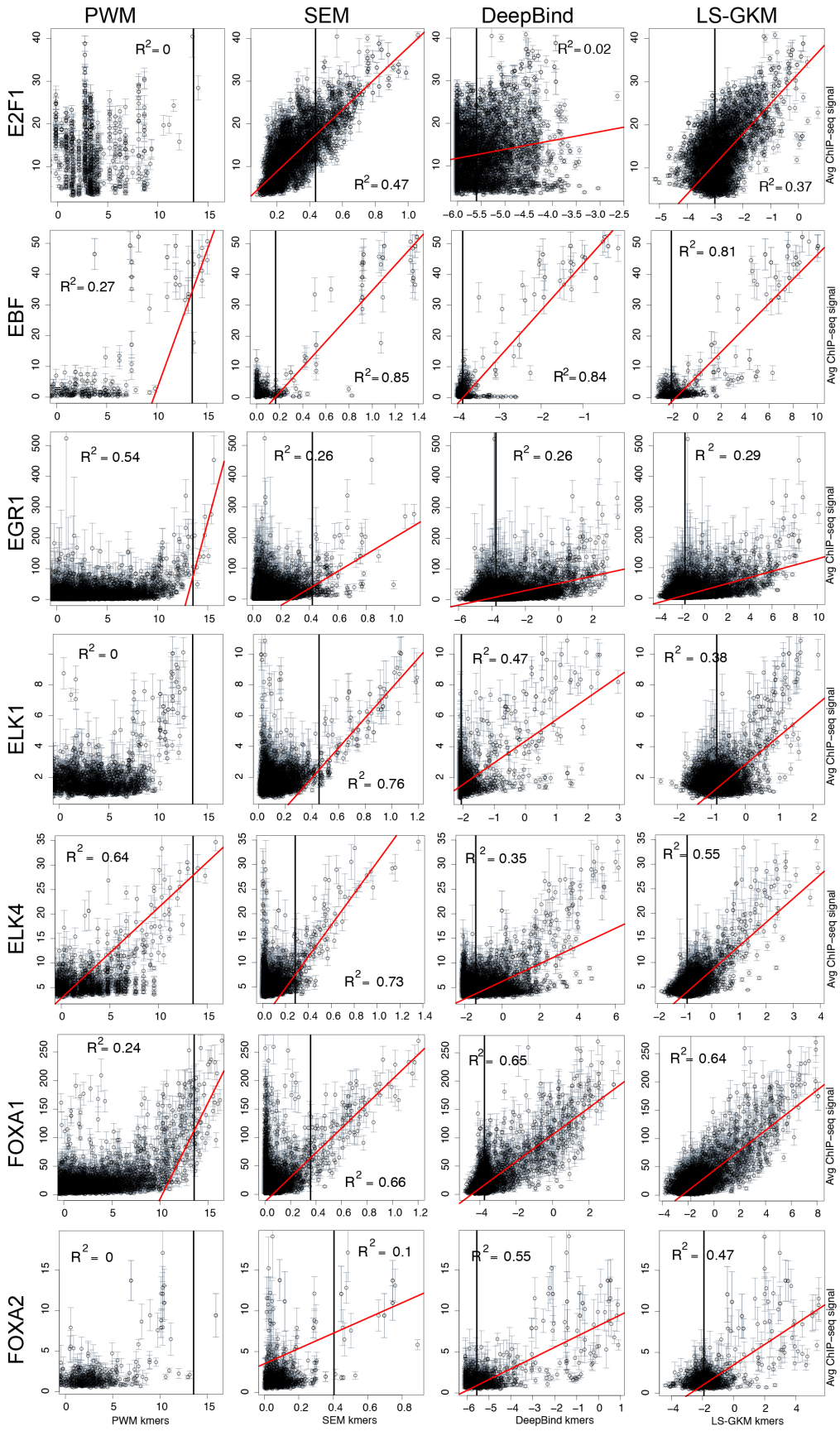


Figure 3.15: SEM scores are correlated with reporter expression changes. Reporter assay expression changes across 5 transcription factors from Kheradpour et al. were compared to SEM score changes [32].

ChIP-seq data than both DeepBind and LS-GKM for 6/13 of the transcription factors tested, and comparably to 3/13 (Figure 3.17). Of the final four transcription factors, two were better predicted by PWMs (EGR1 and MEF2A), HNF4a was poorly predicted by all methods and FOXA2 was best predicted by DeepBind. However, we note that, with some exception, all methods do have good apparent correlation with ChIP-seq data and provide some indication of the effect of variation on TF binding. We would expect transcription factors with binding strongly dependent on sequence outside of the central motif to be better predicted by machine leaning models such as DeepBind and LS-GKM, however for the majority of transcription factors examined here SEMpl predictions based on the central motif were sufficient. This is interesting as it suggests reasonable predictions for transcription factor-binding affinity


PWM
SEM
DeepBind
LS-GKM
E2F1
EBF
EGR1
ELK1
ELK4
FOXA1
FOXA2
R2= 0
R2= 0.47
R2= 0.02
R2= 0.37
R2= 0.27
R2= 0.85
R2= 0.84
R2= 0.81
R2 = 0.54
R2 = 0.26
R2 = 0.26
R2 = 0.29
R2 = 0
R2 = 0.76
R2 = 0.47
R2 = 0.38
R2 = 0.64
R2 = 0.73
R2 = 0.35
R2 = 0.55
R2 = 0.24
R2 = 0.66
R2 = 0.65
R2 = 0.64
R2 = 0
R2 = 0.1
R2 = 0.55
R2 = 0.47
Avg ChIP−seq signal
PWM kmers
SEM kmers
DeepBind kmers
LS-GKM kmers

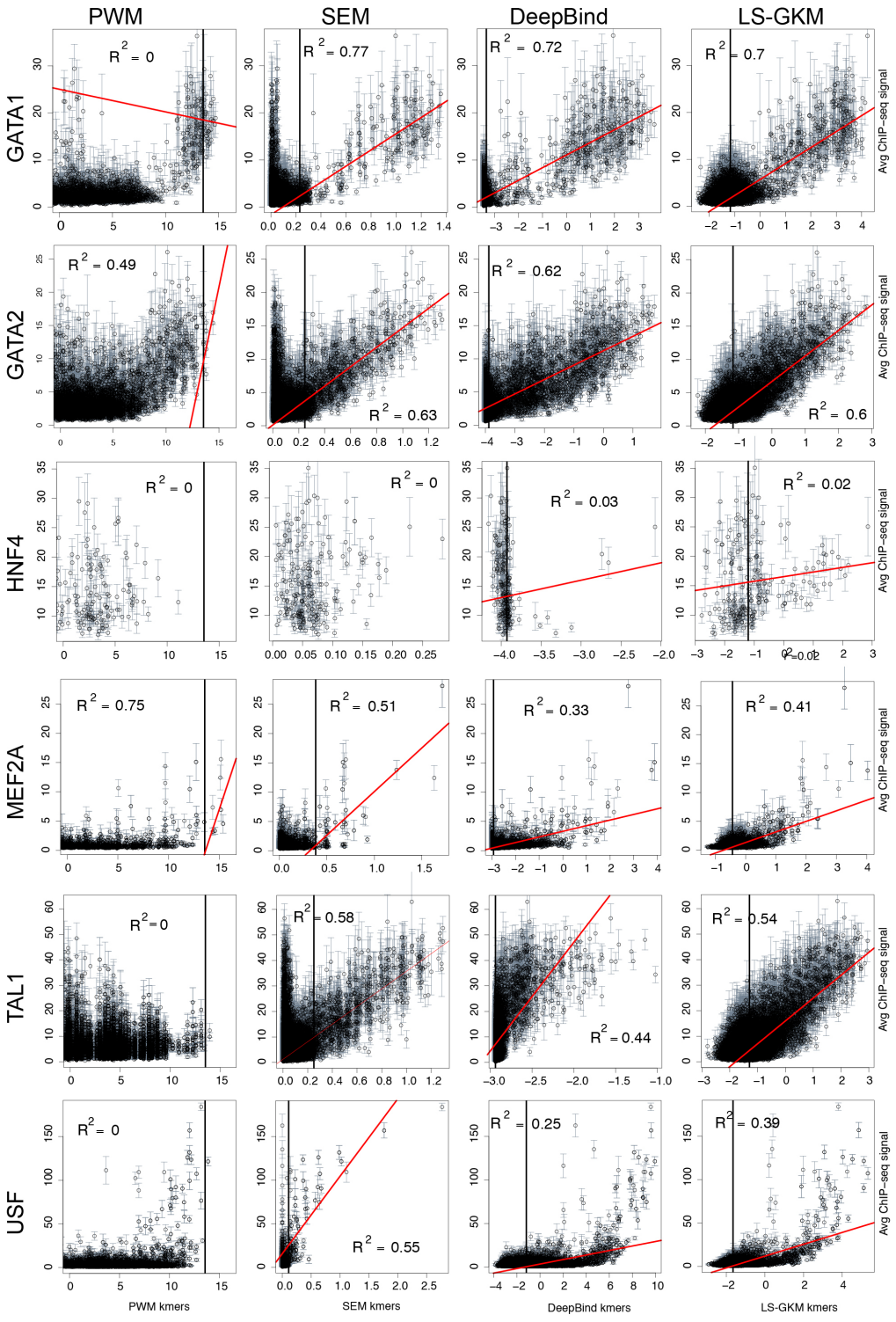


Figure 3.16: Correlations of ChIP-seq data to PWM, SEM, DeepBind, and LS-GKM binding predictions for 13 transcription factors.

changes can be made using a much simpler scoring system, analogous to scoring using a PWM, while avoiding the pitfalls and computational effort required to train a machine learning model. Indeed, by providing pregenerated predictions for many transcription factors, we hope to make using SEMpl as fast and straightforward as possible.

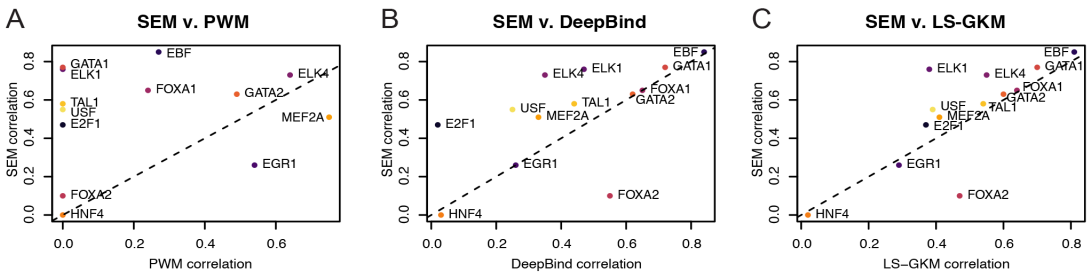


Figure 3.17: Performance comparison of SEMpl to other non-coding SNP prediction methods. Predictions for 13 TFs were generated using PWM (A), SEM, DeepBind (B), and LS-GKM (C) and compared to the average ChIP-seq score for the analogous kmer sequence. Correlations for each transcription factor were then compared across methods. SEMpl produced better or comparable correlations for 9/13 transcription factors tested. PWMs performed better for EGR1 and MEFF2A, and DeepBind performed best for FOXA2. All methods performed poorly for HNF4. The colors/shades of gray of points are unique to each transcription factor.

3.5 Discussion

A deeper understanding of the role non-coding variants play in altering gene expression is critical to fully illustrate the regulatory complexity of our genome and is an important first step toward developing tools for personalized medicine. Approaches such as the IGR method have expanded our ability to use currently available data to predict SNPs that play a regulatory role and have successfully been implemented in multiple studies to link human disease to specific transcription factors and their binding sites. Since its release, the IGR method has been used to successfully identify functional SNPs in TFBSs from GWAS data for breast cancer, atrial fibrillation

and lupus [115, 116, 117]. Functional predictions for these SNPs were experimentally validated, suggesting that the IGR process can be a robust method for functional non-coding GWAS SNP prediction. Unfortunately, this method is not accessible for widespread use. By developing a tool which generalizes the IGR methodology to predict the magnitude and direction of effect of all SNPs within a TFBS, we can identify novel variants associated with human disease in TFBSs genome-wide.

In this article, we introduced SEMpl, a new tool designed to identify putative deleterious mutations in TFBSs. SEMpl predictions reflect known patterns of transcription factor binding while providing additional information about magnitude and direction of predicted change. We demonstrate that SEMpl provides more robustand consistent predictions both on a single variant and a TFBS kmer level than the current standard, PWMs. The method leverages simulation and real data to better model strength of binding rather than a consensus sequence. Additionally, SEMpl scores correlate with known allele-specific binding sites and agree with *in vitro* binding analysis via ChIP qPCR and EMSA as well as previously published variants known to alter transcription factor-binding affinity. Importantly, we found that SEMpl predictions outperform popular machine learning methods for the majority of transcription factors tested.

SEMpl was designed to be easy to use and accessible. In addition to being available as an open source application, precompiled SEM plots for 90 transcription factors from over 200 PWMs are available online. While SEMpl is currently limited to transcription factors with available ChIP-seq and PWM data, we may be able to eliminate the use of PWMs to guide TFBS loci in future versions of our pipe- line, potentially by using overrepresented kmers from the ChIP-seq data, which would reduce bias and expand our list of compatible transcription factors. In addition, we

are working to include additional genomic features, such as DNA methylation which would allow the inclusion of additional bases to SEM plots and a more nuanced understanding of transcription factor binding.

SEMpl's ability to better predict the impact of genomic variation on transcription factor binding has broad implications to the cross- disciplinary study of the regulatory genome. SEMpl has great usability for prioritizing GWAS SNPs for experimental follow-up, in individual studies or through the evaluation of non-coding GWAS catalog SNPs. With the increased need for experimental validations following large-scale genomics studies, we anticipate that annotation tools, such as SEMpl, will be critical in revealing developmental and disease-associated regulatory SNPs.

3.6 Notes & Acknowledgments

This chapter was previously published in *Bioinformatics* (Volume 36, Issue 2) in January, 2020 [118]. The work presented here represents a group effort. I performed all experiments and analysis with the following notable exceptions. This project was originally pioneered by Natalie Ng and Alan P Boyle. Shengcheng Dong completed the allele-specific binding analysis. Robert S Porter generated the CTCF protein used in the EMSA experiments. Cody Morterud and Colten Williams translated the code into C++. Courtney Asman and Jessica A Switzenberg supported all experimental work, notably the mutagenesis of variable regions used in the EMSA experiments.

CHAPTER IV

SEMplMe: A Tool for Integrating DNA Methylation Effects in Transcription Factor Binding Affinity Predictions

4.1 Abstract

Aberrant DNA methylation in transcription factor binding sites has been shown to lead to anomalous gene regulation that is strongly associated with human disease. However, the majority of methylation-sensitive positions within transcription factor binding sites remain unknown. Here we introduce SEMplMe, a computational tool to generate predictions of the effect of methylation on transcription factor binding strength in every position within a transcription factor's motif. SEMplMe uses ChIP-seq and whole genome bisulfite sequencing to predict effects of methylation within binding sites. SEMplMe validates known methylation sensitive and insensitive positions within a binding motif, identifies cell type specific transcription factor binding driven by methylation, and outperforms SELEX-based predictions. These predictions can be used to identify aberrant sites of DNA methylation contributing to human disease. Availability and Implementation: SEMplMe is available from https://github.com/Boyle-Lab/SEMplMe.

4.2 Introduction

DNA methylation is an epigenetic mark as it contributes to changes in the information content of DNA without changing the underlying sequence. The majority of DNA methylation in the human genome occurs at cytosine-phosphate-guanine (CpG) nucleotides. These have long been considered a repressive mark based on early studies of promoters where methylation correlated with transcriptional repression [119]. Methylation at transcription factor binding sites has previously been thought to correlate with the repression of transcription by either disrupting the binding of methylation-sensitive transcription factors or by having no effect on methylation-insensitive transcription factor binding (Figure 4.1A) [120]. However, recent high throughput studies have found that methylation within transcription factor binding sites can lead to increased or decreased transcription factor binding dependent on the position within the motif [121, 122]. Recent work has shown that the strength of the effect of methylation on transcription factor binding affinity varies between nucleotides within a single transcription factor motif. It is vital to determine the specific functional impact of methylation within transcription factor binding sites as aberrant methylation is a hallmark of many human diseases, including cancer, schizophrenia, and autism spectrum disorder [123, 124, 125]. Methods that can better predict these effects on gene transcription can assist in identifying and prioritizing potentially harmful variations.

The effect of methylation on the binding of individual proteins has been studied *in vitro* using protein binding microarrays (PBMs) and newer systematic enrichment of ligands by exponential enrichment (SELEX) based methods [126, 127, 128, 129]. Both PBMs and SELEX rely on proteins binding to DNA fragments *in vitro* and may

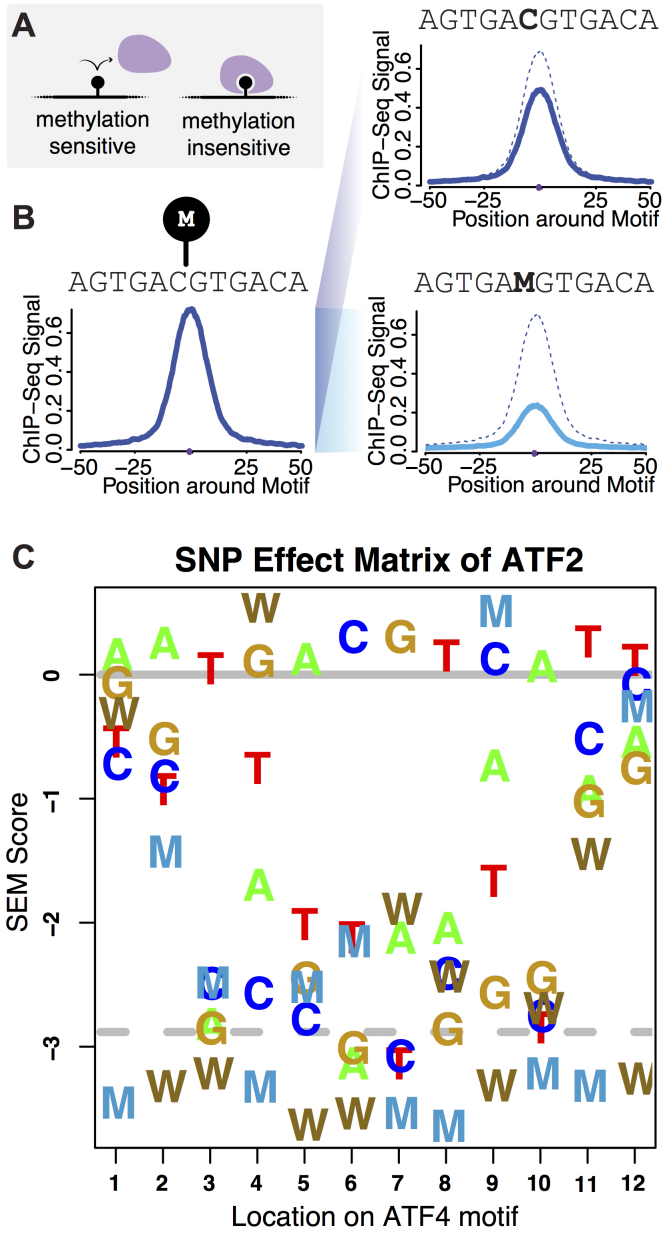


Figure 4.1: SEM pipeline with methylation predicts the effect of methylation on transcription factor binding affinity. A. Traditional model of methylation sensitivity where methylation sensitive transcription factors are unable to bind their site with DNA methylation present, and methylation insensitive transcription factors can bind regardless of the presence of DNA methylation. B. SEMplMe expands on SEMpl output by adding WGBS to divide ChIP-seq signal peaks of C and W into the proportion of their signal affected by DNA methylation using a weighted sum. C. SEMplMe output is displayed as all 6 nucleotides, including methylated C (M), and G opposite to methylated C (W), at every position along the motif. All values are displayed as log 2 and normalized to an endogenous binding baseline set to 0 (dark gray line). A scrambled baseline is also included (dashed gray line).

not recapitulate endogenous binding patterns within the genome. A recent study of methylation sensitivity in 542 human transcription factors using a high throughput SELEX method, methyl-SELEX, found 23% of transcription factors were sensitive to methylation, 34% were enhanced by methylation, and 40% were insensitive to methylation [122]. Computational methods to analyze methyl-SELEX data, such as Methyl-Spec-Seq, provide quantitative information on the magnitude and direction of the predicted effect of methylation on transcription factor binding [129]. Additional SELEX-based studies have also observed differences in methylation sensitivity between different positions in a single motif, and is supported by evidence that some bases within transcription factor binding motifs are more correlated with disease compared to others [130, 87]. However, predictions produced from these methods are limited as they rely on *in vitro* SELEX data and may not reflect binding patterns in a native context. Methods to determine the methylation sensitivity of transcription factors *in vivo* exist, however they are experimentally rigorous, or do not directly estimate methylation consequence on transcription factor binding, and are therefore challenging to use for broad interpretation [121, 131, 132]. A robust method to study the native context of DNA methylation within transcription factor binding sites using *in vivo* data is still needed to more accurately model the role these epigenetic marks play on transcriptional regulation.

To address this need, we have adapted SEMpl, a computational genome-wide transcription factor binding affinity prediction method developed by our lab, to incorporate whole genome bisulfite-seq (WGBS) data. This allows our predictions to include the effects of DNA methylation on binding affinity [118]. SEMpl uses open-source *in vivo* data to generate predictions using transcription factor binding data from ChIP-seq and open chromatin data from DNase-seq for a transcription factor

of interest. The results are displayed as a SNP effect matrix providing predictions for every potential base change in a transcription factor's motif. Our SNP Effect Matrix pipeline with Methylation (SEMplMe) method expands these results by incorporating methylation data from WGBS, generating predictions that encompass the magnitude and direction of change to transcription factor binding for all 4 nucleotide base pairs, and adds two additional nucleotide letters: methylated C (M), and G opposite to a methylated C (W). This new tool provides improved specificity to determine which variants lead to disruption of transcription factor binding by integrating endogenous functional information on methylation states and transcription factor binding, advancing our ability to interrogate and prioritize mutations likely to be associated with human disease.

4.3 Methods

Usage/accessibility

SEMplMe is open source and can be downloaded from https://github.com/Boyle-Lab/SEMplMe. Precomputed SEMplMe plots are available for more than 70 transcription factors.

SNP Effect Matrix pipeline with Methylation

SEMplMe functions as an extension of our previously published method SEMpl. Using the final output of SEMpl as a template, SEMplMe uses whole genome bisulfite sequencing (WGBS) data to evaluate the contribution of DNA methylation on transcription factor binding (Figure 4.1B). WGBS data is gathered for each kmer aligned to the genome containing an in silico SNP. All data shown was generated us-

ing matched cell types for ChIP-seq, DNase, and WGBS data. As the vast majority of sites in WGBS data methylation are not binary, the contribution of the proportion of methylation on binding for C and G SNPs at each position within a motif is calculated. Methylation is calculated for each aligned SNP list using the equation: $\sum_{k=1}^{n} \frac{M*S}{k}$, where M represents the proportion of methylation for an aligned kmer, S represents the ChIP-seq signal for it's alignment, and n represents the total number of kmers in the list. Therefore, the equation: $\sum_{k=1}^{n} 1 - \frac{M*S}{k}$ represents the signal contribution of the non-methylated kmer. Using this method, cytosines are divided into methylated and non-methylated components for each position within the motif of a transcription factor. Following this, all 6 nucleotides are included in a SNP effect matrix at each position along the motif of the transcription factor and plotted for an easy to visualize model of transcription factor binding (Figure 4.1C).

SEMplMe is written in perl and R. In addition to a the matrix file (.me.sem) and the pdf of the visualized sem (_semplot.me.pdf), the output also includes a matrix of standard error (.sterr) and a matrix of total ChIP-seq signal (.me.totals). New alignment and baseline files are also generated for SEMplMe (.me). A quality control file was used, which provides the -log10(P-value) of the average of 100 t-tests from 1000 randomly chosen kmers from the signal files versus the scrambled signal files from SEMpl. A threshold of 3.15 was set to report confidence in a SEM plot, with runs falling under this threshold highlighted in red.

SEMplMe sequence scoring

Scoring a full sequence with SEMplMe can be done in the same manner as PWMs or SEMpl, where the log2 score analogous to the nucleotide of interest at each position is added to reflect the predicted binding score of the sequence. This allows predic-

tions to be made for motifs carrying more than one variant.

EMSA

Kd values for CEBPB and ATF4 were calculated from a previously published EMSA reaction by densiometric scanning by ImageJ and the Excel Solver Package [112, 127, 113]. All EMSA scores are represented as a ratio to the unmethylated control.

Correlation with ChIP-seq data

All kmers likely to bind CTCF were recovered from the final iteration of SEMpl. For each kmer with at least 50 occurrences, the average ChIP-seq signal and standard error were calculated. Correlation cutoffs for SEMplMe were defined as the scrambled baseline for the final iteration of SEMpl.

4.4 Results

SEMplMe provides quantitative predictions based on *in vivo* measures of binding affinity

SEMplMe integrates endogenous functional data encompassing transcription factor binding, open chromatin, and DNA methylation to provide a quantitative prediction of the effect of methylation on transcription factor binding affinity at every position within a binding motif. By including measures of DNA methylation, SEMplMe is able to calculate the relative average transcription factor binding affinity of methylated genomic sequence by using a weighted sum of ChIP-seq signal and the proportion of methylation at the site from WGBS (Figure 4.1B). Averaging this signal genome-wide for methylated and unmethylated sequence separately allows for

the generation of a quantitative prediction matrix of the effect methylation has on transcription factor binding affinity (Figure 4.1C). SEMplMe represents an advancement over currently existing methods as its predictions are generated from *in vivo* functional data, it is generally accessible without additional experimental work, and the resulting matrix is both quantitative for a single position and across an entire motif.

SEMplMe recapitulates differences in methylation sensitivity between transcription factors

Transcription factor differences in methylation sensitivity were examined by calculating the absolute difference between methylated and unmethylated bases at each position within SEMplMe matrices for methylation sensitive and insensitive transcription factors. Methylation sensitive transcription factors examined here include CREB, cMYC, USF, NFkB, E2F, MYC, and ZFX [133, 120, 134, 129]. Methylation insensitive transcription factors examined here include SP1, REST, CEBPa, FOXA1, RXRA, and ARNT2 [135, 126, 133, 120, 136, 129]. As expected, transcription factors previously associated with methylation sensitivity show a larger average difference in SEM scores between C and M, and G and W nucleotides compared to transcription factors previously defined as insensitive (Figure 4.2). This suggests that prior definitions of methylation sensitivity and insensitivity may reflect general trends of transcription factor methylation sensitivity. However, it remains unclear if this trend is driven from methylation sensitivity across an entire motif, or typically driven by a single position.

DNA methylation drives cell type specific transcription factor binding

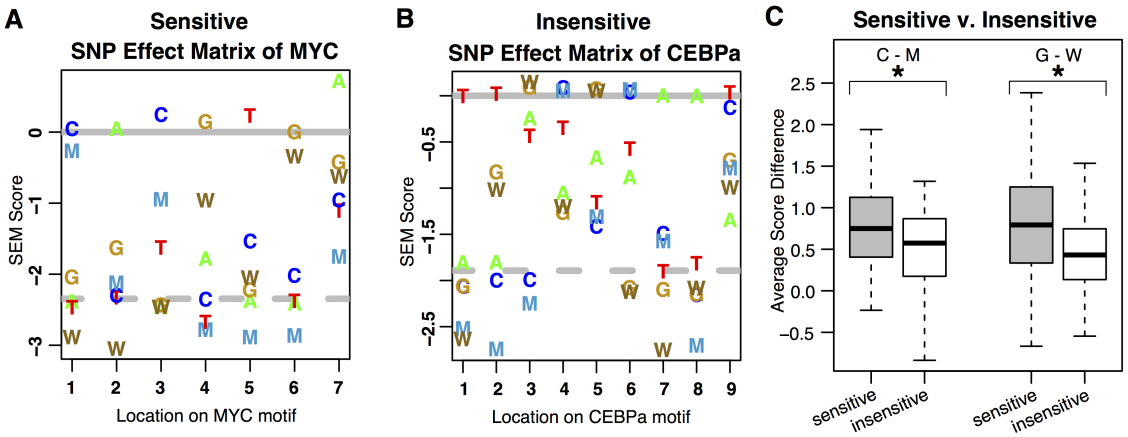


Figure 4.2: SEMplMe confirms differences in methylated SEM scores for sensitive versus insensitive transcription factors. A. Known methylation sensitive transcription factor MYC shows a large difference between methylated and unmethylated nucleotides at most positions. B. Known methylation insensitive transcription factor CEBPa shows very little difference between methylated and unmethylated nucleotides at most positions. For some positions (i.e. position 5), a small increase in binding is predicted for a methylated cytosine. C. Transcription factors previously annotated as methylation sensitive and insensitive show a significant difference in methylated (M/W) and non-methylated (C/G) SEM scores (T-test C-M P–value = 0.007 and G-W P-value = $1.32*10^7$). Error bars represent standard deviation.

DNA methylation is hypothesized to contribute to cell type specific transcription factor binding by altering the availability of DNA sequence. In support of this, the input cell type was found to influence the output of SEMplMe for some transcription factors. One example, JUN, shows high correlation of SEMplMe outputs for methylated sites (MW) between H1-hESC and K562 cell lines ($R^2 = 0.91$), and a reduced correlation to HepG2 ($R^2 => 0.43$) (Figure 4.3). This is supported by MethMotif data, in which JUN shows many more methylated binding sites, most of which fall into a mid- to highly-methylated state in HepG2, as opposed to comparatively few overlapping methylated sites in K562 and H1-hESC [132]. This pattern of reduced correlation was not observed when looking across the entire SEMplMe output, suggesting methylated sites are driving this difference (Figure 4.4). Of note, this pattern is not seen for another transcription factor, CEBPB, where the SEMplMe output for methylated sites is highly correlated between all cell types examined (K562, IMR-

90, HepG2, and GM12878), suggesting that not all transcription factors are subject to cell type specificity due to methylation differences (Figure 4.5). Interestingly, SEMpl data without methylation appears to be primarily cell type agnostic, providing evidence that methylation plays a meaningful role in cell type specificity for some transcription factors [118].

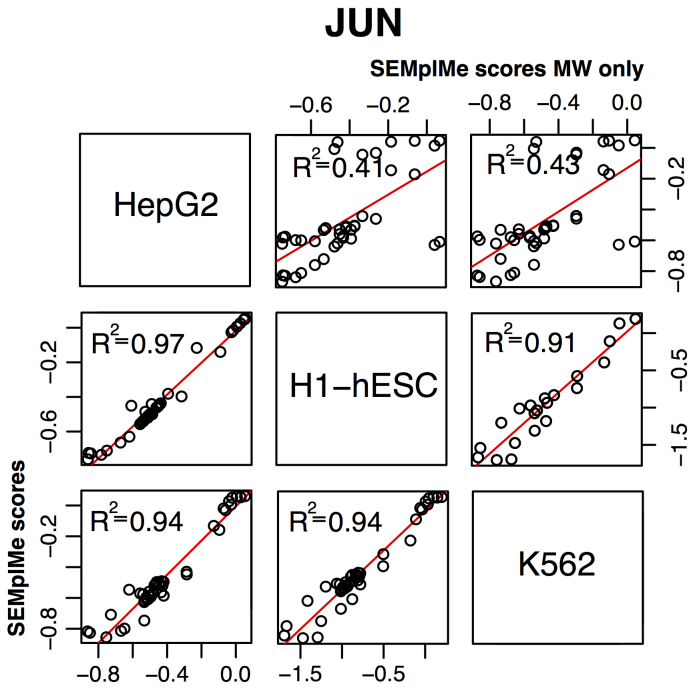


Figure 4.3: SEMplMe output for JUN varies between cell types. SEM plots vary more between cell types when only considering methylated sites (top right) than methylated and unmethylated sites (bottom left). This suggests methylation plays a key role in the cell type specificity of the transcription factor JUN.

SEMplMe validation using *in vitro* measures of transcription factor binding affinity

To evaluate SEMplMe on a metric external to ChIP-seq data, our predictions were compared to PBM data, which has been used by previous studies to examine the affinity of individual transcription factors to potential target sequence *in vitro* [126, 127, 128]. SEMplMe predictions were compared to microarray Z-scores data from

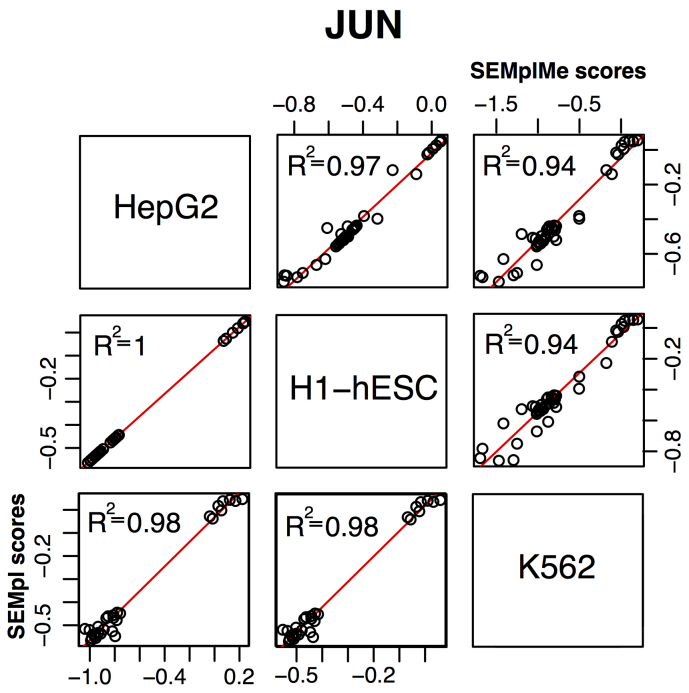


Figure 4.4: SEMplMe output between cell types versus SEMpl without methylation. SEM plots show more variance between cell types for SEMplMe (top right) than SEMpl without methylation (bottom left).

PBMs, which represent transcription factor binding affinity to methylated or unmethylated DNA sequence. A high level of agreement was observed between SEMplMe predictions and previously published PBM data across 8 transcription factors (Figure 4.6A)(R^2=0.67)(CEBPA, CEBPB, CEBPD, CREB1, ATF4, JUN, JUND, CEBPG) [128]. This agreement is reduced when using SEMpl scores without methylation (R^2=0.56), suggesting that the inclusion of methylation in our model improves scores for methylated sequences (Figure 4.6B). Discrepancies between SEM predictions and PBM data can be attributed to differences in *in vivo* versus *in vitro* methods of generation.

To further functionally validate SEMplMe, data from *in vitro* electrophoretic mobility shift assays (EMSAs) were utilized to examine our predictions. Previously published EMSA data was evaluated for two transcription factors, ATF4(CREB)

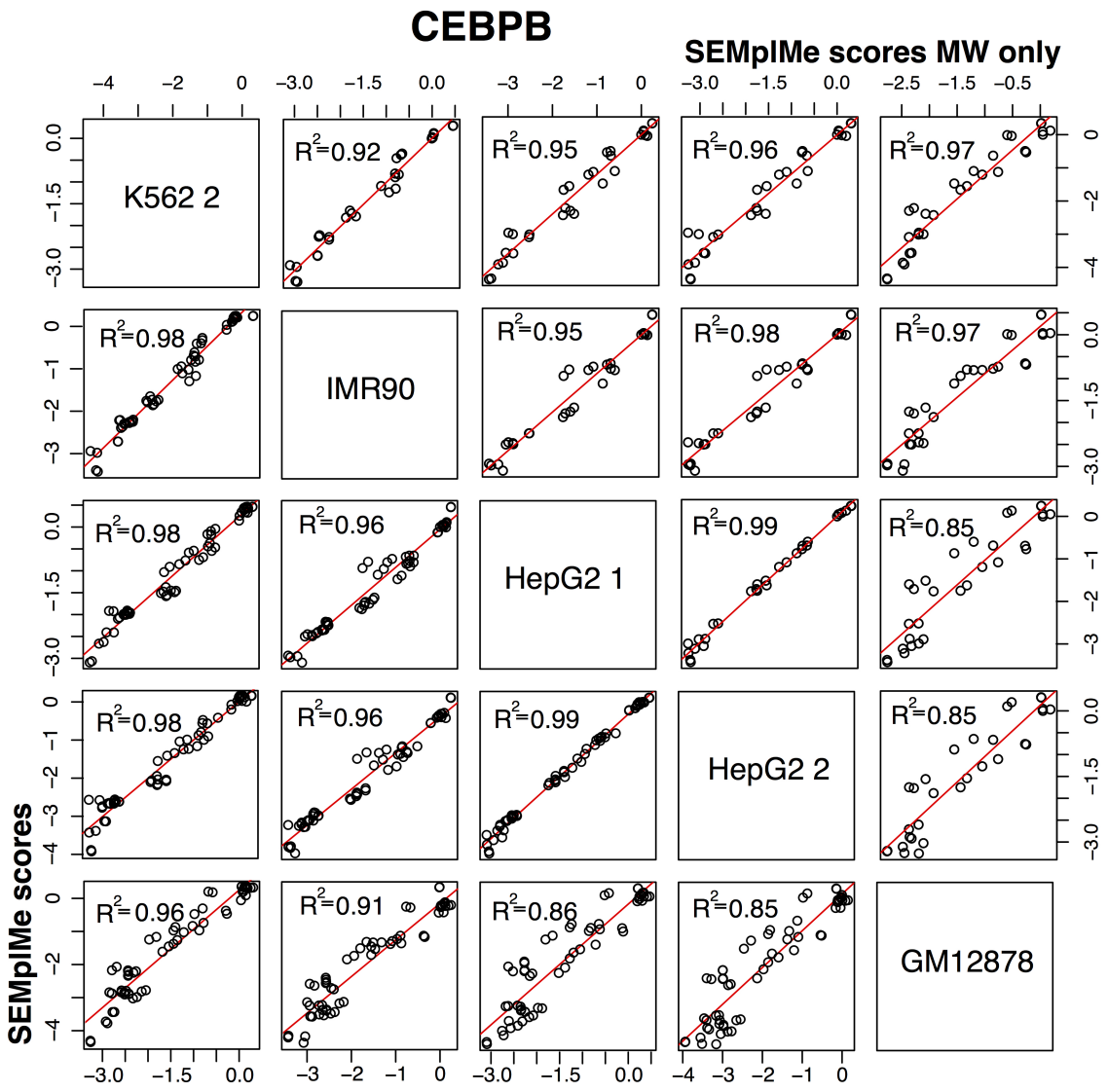


Figure 4.5: CEBPB SEM output between cell types. SEM plots show little variance between cell types when considering only methylated sites (top right) or both methylated and unmethylated sites (bottom left) for CEBPB. This suggests methylation does not play a large role in cell type specificity for CEBPB.

and CEBPB. This measure of *in vitro* binding showed marginal agreement with our predictions (R^2=0.65)(Figure 4.6C)[127]. This observed low agreement is driven entirely by CEBPB which has relatively low correlation with our predictions (R^2=0.17), as opposed to ATF4 (R^2=0.92). CEBPB has been reported to preferentially bind to methylated sequence, thus the discrepancy in predictions has previously been

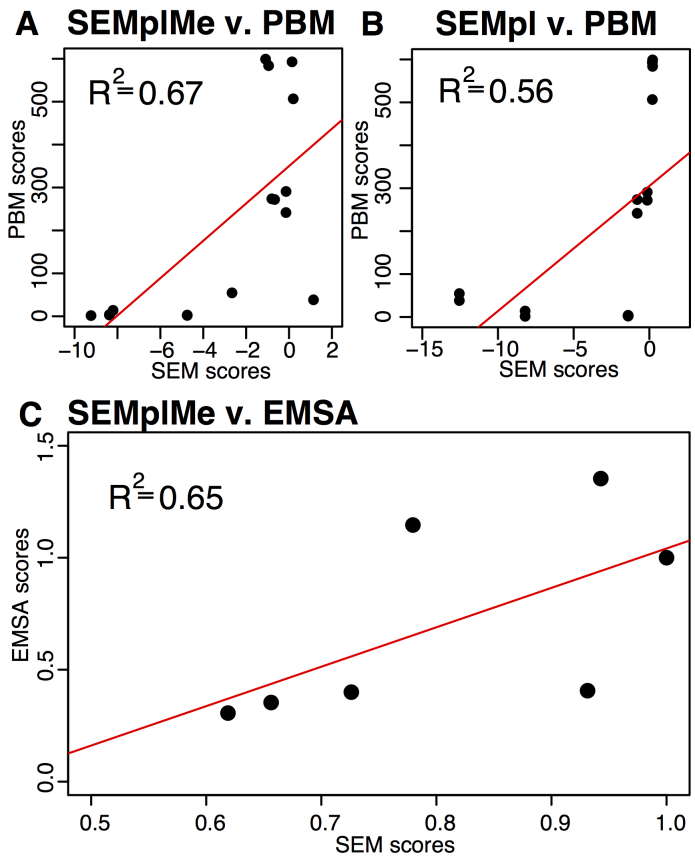
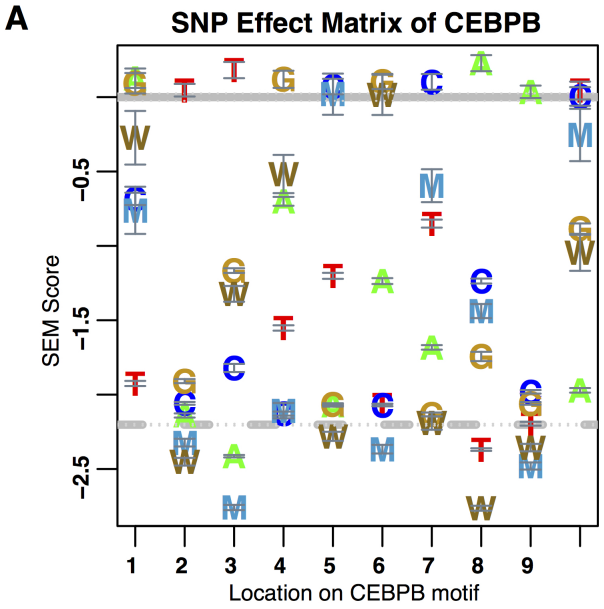


Figure 4.6: SEMplMe predictions agree with in vitro experimental methods. A. SEMplMe agrees with previously published protein binding microarray (PBM) data of methylated and unmethylated binding sites for 8 transcription factors ($R^2 = 0.67$) [128]. B. SEMpl shows a reduced correlation with PBM data compared to SEMplMe ($R^2 = 0.56$), suggesting the addition of methylation data improves methylated sequence predictions. C. SEMplMe predictions correlate with previously published electrophoretic mobility shift assay (EMSA) data for methylated, hemi-methylated, and unmethylated binding sites for ATF4 and CEBPB ($R^2 = 0.65$) [127].

thought to be a result of limited genome methylated sequence availability, a necessity for calculating more accurate predictions in SEMplMe [127]. SEMplMe identified comparatively few methylated sites throughout the genome, leading to a much higher standard deviation for the effect of methylated sites (Supplementary Figure 3). This unavailability of methylated sites is consistent with previous data showing methylated CEBPB motifs to bind well *in vitro*, but poorly *in vivo* [137].



B

CEBPB Loci Totals

	A	C	G	T	M	W
1	1856	2499	2154	2068	134	152
2	4244	6007	4147	4987	332	284
3	2519	2799	6003	3708	296	644
4	3551	3905	3289	3716	197	255
5	6276	3588	6000	5342	466	420
6	5517	6029	3762	6421	426	487
7	4406	3687	4231	4019	301	224
8	3812	6263	2539	2573	716	292
9	5052	4767	6629	4514	315	350
10	2450	2372	2954	2060	167	170

Figure 4.7: Total number of kmers for each nucleotide in the SEM of CEBPB. A. SEM plot of CEBPB with error bars representing standard deviation. B. Counts of mapped kmers in the genome for each nucleotide at each position. These counts are inversely proportional to the standard deviation seen in the SEM.

SEMplMe predictions are consistent with previous findings for CTCF

CTCF is a well studied transcription factor previously shown to be methylation sensitive [138, 139]. CTCF binding predictions using SEMplMe found the majority of positions to be methylation sensitive for both M and W. Notably, a handful of sites had methylated sequence scores at or slightly above their unmethylated counterpart, and likely represent methylation insensitive positions. These results are consistent with CTCF's role as a methylation sensitive transcription factor. As CTCF is widely used in research studies, its binding to sites containing methylated positions

within its motif have been previously surveyed by a variety of methodologies, including qualitative EMSA, observation of binding following demethylation of cells, and SELEX-based methods [138, 121, 140, 110]. When SEMplMe results were compared to measures of binding at individual positions within the CTCF motif, a general agreement was observed for the direction of binding for all positions predicted to decrease binding affinity (Figure 4.8). Though the majority of sites identified by previous studies within the CTCF motif were found to be overwhelmingly methylation sensitive, two sites were predicted to lead to increased binding affinity when methylated. Though SEMplMe did not identify these positions, one site overlaps a SEMplMe position consistent with methylation insensitivity, and the other was found to not significantly increase binding by all prior studies [121]. Overall, our predictions are consistent with previous studies of CTCF binding direction.

Correlation between the entirety of the CTCF matrices generated by SEMplMe and the recently published Methyl-Spec-seq method, which uses *in vitro* SELEX to predict methylation effects on transcription factor binding affinity, was assayed (R^2=0.56) (Figure 4.9A)[110]. SEMplMe outperformed Methyl-Spec-seq by performance comparison when comparing scores across entire kmers to their average ChIP-seq signal (SEMplMe R^2=0.25, Methyl-Spec-seq R^2=0.04) (Figure 4.9B&C). The kmer set used is associated with active CTCF binding and includes both methylated and unmethylated sequences. This provides further evidence that predictions of change to transcription factor binding affinity perform better when generated from *in vivo* data, rather than *in vitro* data such as from SELEX methods.

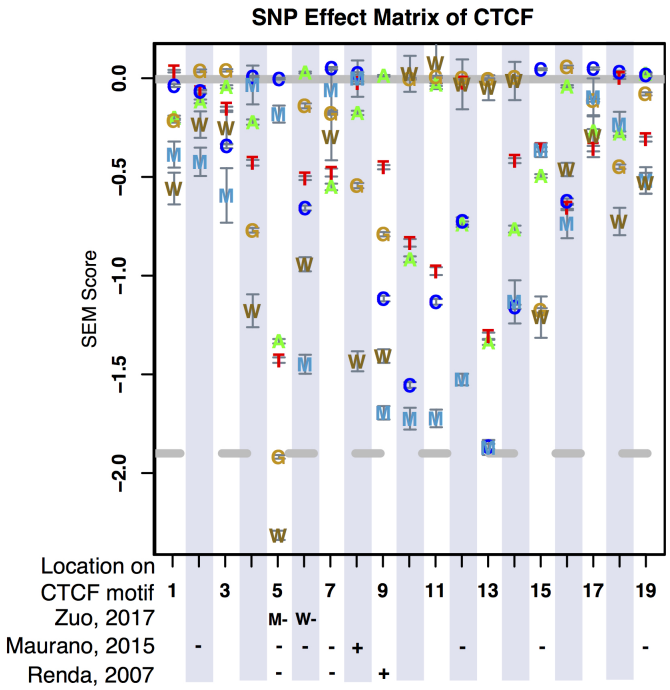


Figure 4.8: SEMplMe predictions agree with previously published predictions and experimental measures of CTCF binding to methylated sequence. Signs (+/-) found below the SEM plot represent the reduction or increase in binding affinity reported by previous studies at the analogous position. All signs shown without an M or W represent a methylated cytosine (M). Error bars represent standard deviation. [110, 121, 140]

4.5 Discussion

DNA methylation is a key epigenetic mark known to act in a regulatory capacity, allowing transcription factors to bind in a cell-type specific manner. Counter to the idea that all methylation is able to disrupt transcription factor binding, recent studies have revealed that certain methylated loci impact binding more than others. Predicting the locations of these methylation sensitive loci and quantifying the effect of methylation on transcription factor binding affinity is challenging. Here we introduce an expansion to our previously released software SEMpl, called SEMplMe, which integrates predictions of the effect of cytosine methylation on transcription factor binding affinity based on WGBS data. These predictions agree with *in vitro* data of transcription factor binding, are cell-type specific, and show a general agreement

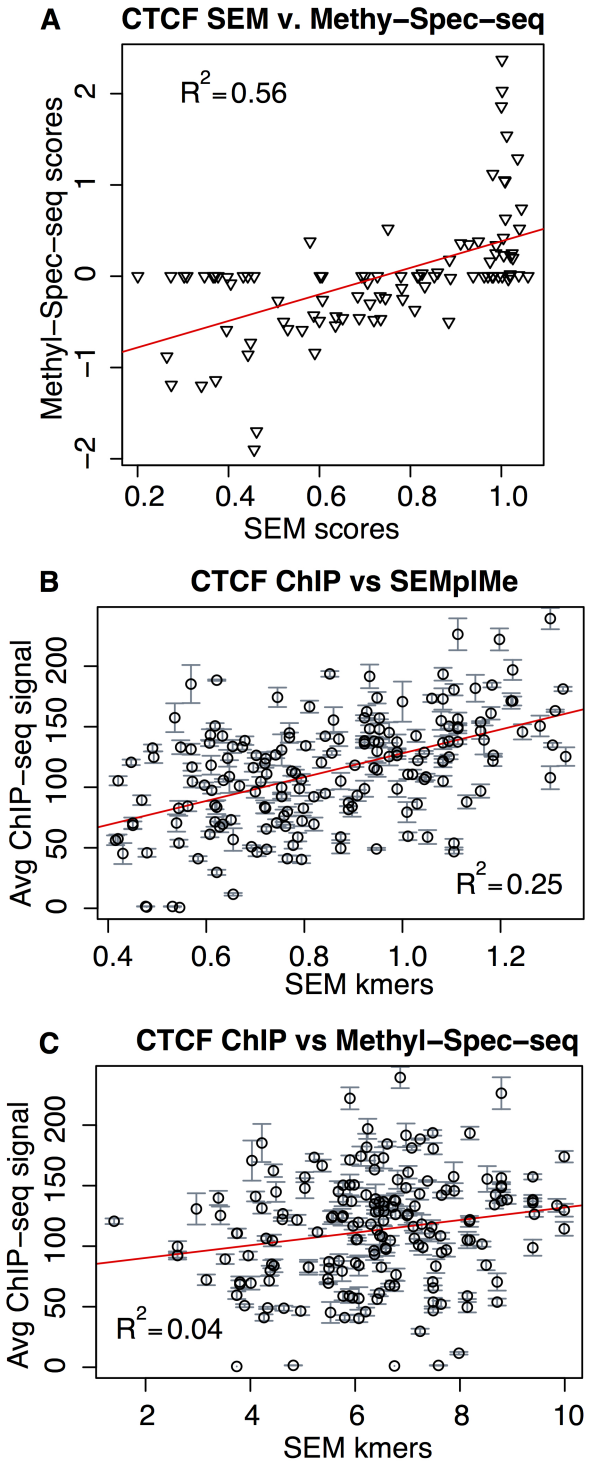


Figure 4.9: SEMplMe has higher correlation with *in vivo* CTCF binding than Methyl-Spec-seq. A. Correlation of CTCF matrices between SEMplMe and Methyl-Spec-seq show a modest agreement (R^2=0.56). B&C. SEMplMe outperforms Methyl-Spec-seq when comparing to CTCF ChIP-seq scores for whole kmers, including methylated sites (SEMplMe R^2=0.25, Methyl-Spec-seq R^2=0.04).

with data from transcription factors previously annotated as methylation sensitive and insensitive.

SEMplMe is poised to advance our understanding of the effects of methylation on transcription factor binding affinity through its generation of quantitative predictions using *in vivo* functional data. SEMplMe will both improve our ability to predict putative disease loci affected by aberrant DNA methylation, and increase predictions of transcription factor binding affinity in general [133]. This is expected to hold true regardless of whether reduced methylation in a transcription factor's motif contributes to its binding, or is caused by its binding [141]. The nucleotide W was included to capture not just position dependent, but stand dependent methylation, as strand specificity due to hemi-methylation has previously been found to influence transcription factor binding [110]. This is likely driven by changes in DNA structure.

SEMplMe has similar limitations to its predecessor SEMpl, such as a dependence on available ChIP-seq, DNase-seq, and WGBS data. It is further restricted by the limited number of methylated sites in the genome available for use in generating models of binding. In instances where few sequence specific sites also contain methylation, our measure of standard deviation increases considerably. Though the low confidence in these sites can be visualized by error bars, predictions of methylation at these loci are limited. Additionally, cell type should be carefully considered before running SEMplMe for optimal predictions as cell type specificity contributes to the final SEMplMe plot, and methylation sensitivity has been previously found to be paralog specific [130].

The inclusion of CpG methylation provides additional information to help fully understand context-specific transcription factor binding. However, the addition of

more nuanced molecular mechanisms that contribute to transcription factor binding are likely to further improve our predictions. This includes additional types of DNA methylation, such as hydroxymethylation and nonCpG methylation, as well as measures of structural changes to the genome [130, 46, 142, 128, 143].

The improved predictions provided by SEMplMe will contribute to a better understanding of the key positions within transcription factor binding sites affected by DNA methylation. This advancement is central to improving our ability to prioritize mutations associated with aberrant methylation contributing to human disease.

4.6 Notes & Acknowledgments

This chapter was previously published to *bioRxiv* in August, 2020 [144].

CHAPTER V

The Inducible *lac* Operator-repressor System is Functional in Zebrafish Cells

5.1 Abstract

Zebrafish are a foundational model organism for studying the spatio-temporal activity of genes and their regulatory sequences. A variety of approaches are currently available for editing genes and modifying gene expression in zebrafish including RNAi, Cre/lox, and CRISPR-Cas9. However, the *lac* operator-repressor system, a component of the *E. Coli lac* operon which has been adapted for use in many other species and is a valuable, flexible tool for studying the inducible modulation of gene expression, has not previously been tested in zebrafish. Here we demonstrate that the *lac* operator-repressor system robustly decreases expression of firefly luciferase in cultured zebrafish fibroblast cells. Our work establishes the *lac* operator-repressor system as a promising tool for the manipulation of gene expression in whole zebrafish. Our results lay the groundwork for the development of *lac*-based reporter assays in zebrafish, and add to the tools available for investigating dynamic gene expression in embryogenesis. We believe that this work will catalyze the development of new reporter assay systems to investigate uncharacterized regulatory elements and their cell-type specific activities.

5.2 Background

Experimental approaches for the study of transcriptional regulation by cis-regulatory elements *in vivo* require methods for both genetically modifying cells or organisms and for measuring expression levels of specific genes. Zebrafish (*Danio rerio*) is an ideal model organism for investigating the spatio-temporal-specific regulation of gene expression throughout the developing embryo as it satisfies the requirements for ease of genetic manipulation and expression readout. Microinjection of DNA into fertilized embryos allows for simple and effective delivery of genome-modification tools, such as Tol2 transposons, that mediate genomic integration of constructed expression cassettes. Additionally, the transparency of zebrafish embryos facilitates the observation of fluorescent signal from reporter genes within live cells and tissue. Due to its benefits as a model organism, many technologies for studying gene function have been developed in zebrafish, including Cre/Lox [145], tamoxifen-inducible Cre [146], the Tet-On system [147], RNAi [148, 149], and more recently, CRISPR based-methods [150]. However, the use of the *lac* operator-repressor system, a tool which functions transiently in a native context with minimal disruption of local regulation compared to many of the aforementioned methods, has yet to be demonstrated in zebrafish.

The *lac* operator-repressor system is an inducible repression system established from studies of the *lac* operon in Escherichia coli (*E. Coli*) that regulates lactose transport and metabolism [151]. The Lac repressor (LacI) binds specifically to a *lac* operator sequence (*lacO*), inhibiting the *lac* promoter and *lac* operon expression through steric hindrance [152]. Addition of the allosteric inhibitor Isopropyl β-d-1-thiogalactopyranoside (IPTG) to cells frees the *lac* operon to express its associated

gene by inhibiting the binding of LacI to *lacO* sequences. The use of IPTG with the *lac* operator-repressor allows for inducible reversal of transcriptional repression.

Since its discovery in prokaryotes, the *lac* operator-repressor system has been modified for use in eukaryotic organisms to study the regulation of gene transcription [152, 153, 154, 155]. Experiments in mammalian cell lines from mouse, monkey, and human [152, 156, 154, 155], as well as in whole mouse [157], demonstrate the utility of the *lac* operator-repressor system. It has also successfully been applied in cell lines and whole organisms of the amphibian axolotl, suggesting that this system can be utilized in a wide range of organisms [158]. Modifications to the *lac* operator-repressor system has allowed for constitutive, ubiquitous expression [152, 159, 160], visually assessed output [156, 155], and the ability to study both gene repression and activation [161, 162], emphasizing its flexibility for studying gene expression dynamics. The ability of IPTG to relieve repression in the *lac* system makes it a more adaptable tool for studying the temporal dynamics of gene expression, compared to constitutively active or repressed reporter gene systems.

In this paper, we provide evidence that the *lac* operator-repressor system can function in the zebrafish fibroblast cell line PAC2, adding a versatile new tool for the study of zebrafish genetics and transcriptional regulation. The results in a zebrafish cell line support the potential functionality of the *lac* operator-repressor system to function in whole zebrafish. In addition, we demonstrate that the CMV-SV40 enhancer-promoter produces strong, widespread reporter gene expression in both a zebrafish cell line and embryos. This enhancer-promoter combination provides a flexible, non-tissue specific expression module for zebrafish to aid in reporter gene detection at a cellular level.

5.3 Results

The CMV-SV40 enhancer-promoter shows widespread expression in zebrafish

In order to promote the repression of a reporter gene in our assay, we sought to increase LacI expression in transfected cells by including a strong enhancer-promoter driving LacI. The CMV enhancer and promoter are frequently used in reporter vector construction across a wide range of studies due to their strong and constitutive promotion of gene expression. This includes zebrafish, where the CMV enhancer-promoter has previously been shown to have strong tissue-specific expression [163]. However, recent studies have demonstrated that tissue-specific promoter function of non-CMV promoters may be lost when paired with the CMV enhancer [164]. With the goal of identifying enhancer-promoter pairs driving non-tissue specific gene expression in whole zebrafish we examined CMV enhancer activity paired with a non-CMV promoter, the SV40 minimal promoter. The function of the CMV-SV40 enhancer-promoter was first validated in PAC2 cells, by inserting them upstream of luciferase in a pGL3 plasmid. Relative luciferase output of the CMV-enhanced SV40 pGL3 plasmid was compared to a pGL3 plasmid containing only a minimal SV40 promoter. The CMV-SV40 enhancer-promoter was able to drive a 24-fold increase in luciferase expression compared to the promoter-only control, suggesting that the CMV-SV40 enhancer-promoter is able to function as a strong enhancer-promoter combination in PAC2 zebrafish fibroblast cells (Figure 5.1A).

To determine if the CMV-SV40 enhancer-promoter was able to enhance reporter expression in whole zebrafish, the Tol2 transposon system was utilized to integrate eGFP-expressing test plasmids into zebrafish embryos. As in PAC2 cells, two con-

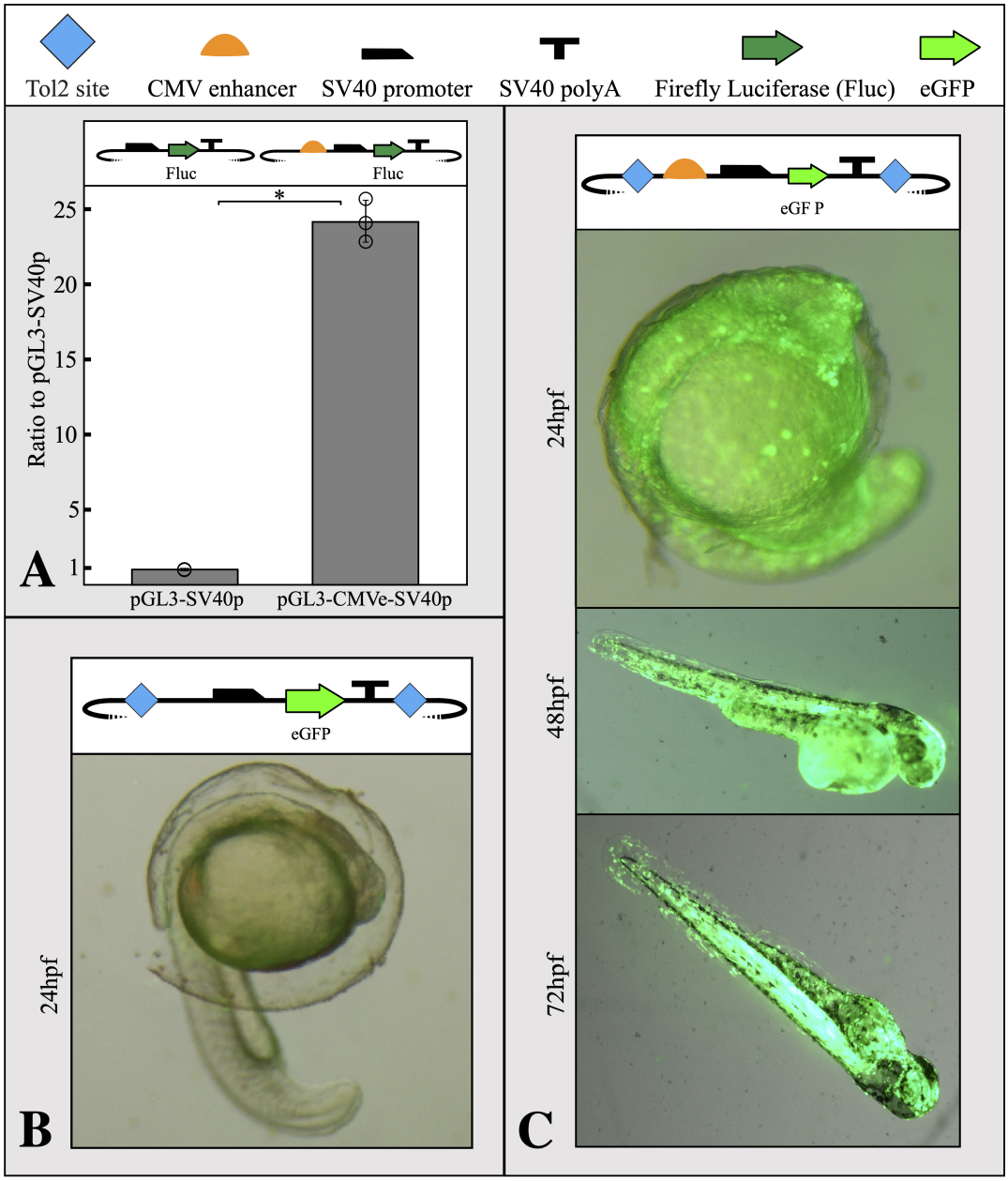


Figure 5.1: A) CMV-SV40 enhancer-promoter shows a 24-fold increase in luciferase activity in PAC2 zebrafish fibroblast cells when transfected with a CMV-SV40 enhancer-promoter driving luciferase compared to a SV40-only plasmid. Error bars represent standard deviation of 3 biological replicates. Statistical significance determined using a stand 2-sided T-test * P-score < 0.01. Points represent values for all 3 replicates in each condition. B) SV40 promoter-only plasmids did not result in observable eGFP expression. C) CMV-SV40 enhancer-promoter driving eGFP shows non-tissue specific expression in zebrafish up to 72 hours post-fertilization. All plasmid components for each transfection design are detailed as symbols at the top of the figure. The TSS begins where the SV40 promoter begins to slope downward.

structs were evaluated; one containing a CMV-SV40 enhancer-promoter upstream of an eGFP reporter gene, and one with only a minimal SV40 promoter. While no detectable level of eGFP activity in the promotor-only control was observed (Figure 5.1B), composite brightfield and GFP images of 24, 48 and 72 hours post-fertilization embryos injected with the CMV-SV40 enhancer-promoter construct show non-tissue specific eGFP expression (Figure 5.1C).

The *lac* operator-repressor system is functional in the PAC2 zebrafish cell line

To test the functionality of the *lac* operator-repressor system, a repressible reporter plasmid containing 6 *lac* operators in the 5'UTR of the firefly luciferase gene and a LacI-expressing plasmid were co-transfected into PAC2 cells. When a plasmid expressing a non-functional LacI (NFLacI) gene was co-transfected, no repression was observed (Figure 5.2), whereas a plasmid expressing CMV enhancer-driven levels of LacI resulted in about 65% repression. An intermediate level of repression ($\sim$40%) was observed when LacI was expressed from a plasmid containing only a SV40 minimal promoter, indicating that the extent of repression correlates with LacI levels in the cell. Addition of IPTG to the cells resulted in full relief of repression in all cases. This indicates that LacI is responsible for repression of luciferase expression in these cells.

To compare PAC2 repression levels to previously published data [165] and test broad functionality within different cell lines, we replicated the LacI experiment in the K562 human cell line. When both cell types were co-transfected with the same plasmid mixture, the performance of the lac operator-repressor system was nearly identical in PAC2 and K562 cells (Figure 5.3). Both PAC2 and K562 cells

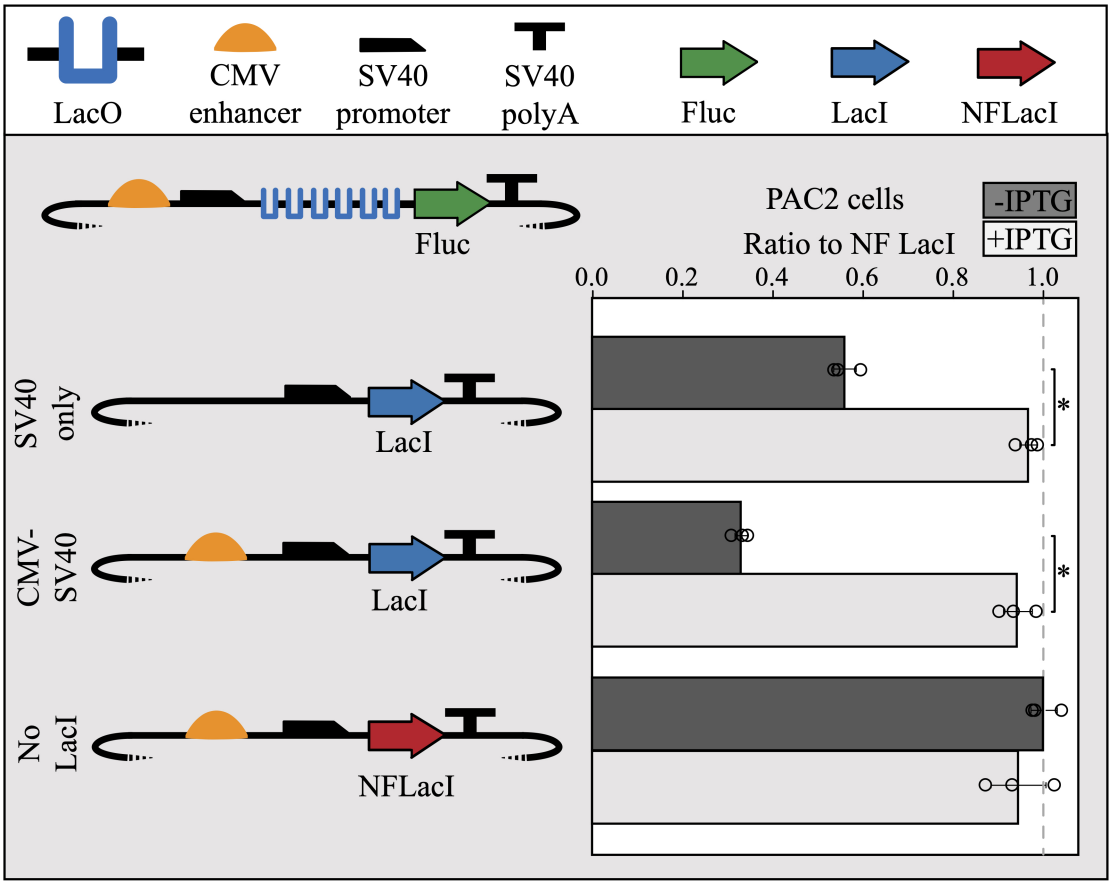


Figure 5.2: Co-transfection of LacI-expressing modules with repressible reporter modules result in LacI-mediated repression in PAC2 cells. SV40 promoter-only driven expression of LacI shows moderate repression (40%) and CMV-SV40 enhancer-promoter driven expression LacI shows high repression (70%) of a repressible module containing 6x *lacO* sites. Expression of non-functional LacI (NFLacI, frameshift mutant) shows no repression and all modules showed maximal reporter expression in the presence of 1mM IPTG. Error bars represent standard deviation of replicates (n=3). Points represent values for all 3 replicates in each condition. The dashed line shows the NF LacI IPTG- negative control. The TSS begins where the SV40 promoter begins to slope downward. Statistical significance determined using a stand 2-sided T-test * P-score < 0.001.

showed around 60-65% repression when co-transfected with a molar equivalent of CMV enhancer-driven LacI containing plasmid (∼400ng), and roughly 10-20% repression when co-transfected with a similar molar equivalent of SV40 promoter-only driven LacI-expressing plasmid. These results demonstrate that the *lac* operator-repressor system functions in PAC2 cells at a comparable level to human K562 cells.

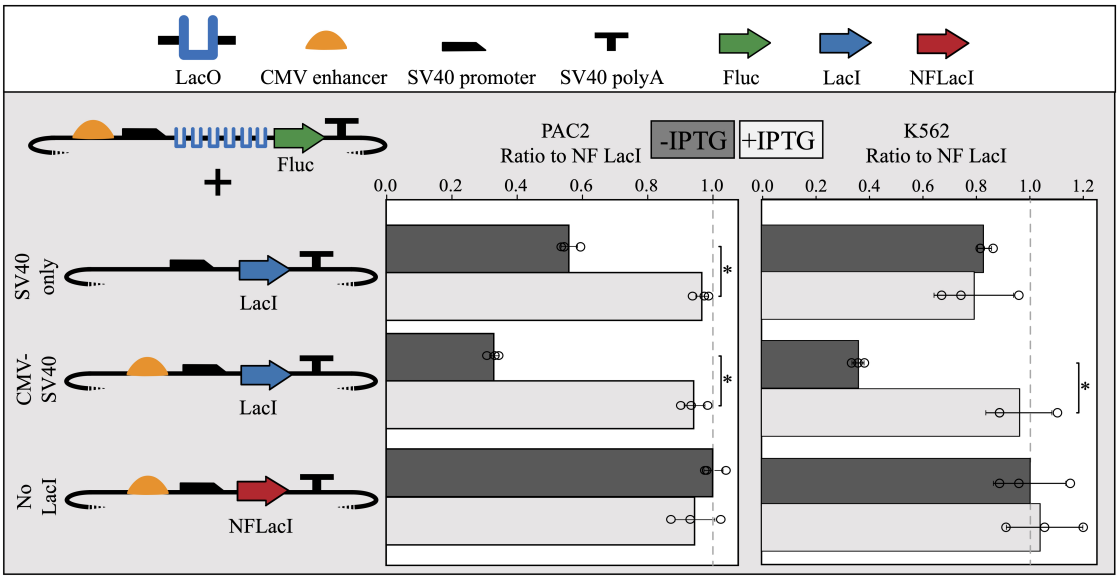


Figure 5.3: The *lac* operator-repressor system performs similarly in both human and zebrafish cell lines. K562 and PAC2 cells transfected with the same *lac* operator-repressor plasmid mixtures result in similar repression profiles. The promoter-only LacI-expressing plasmids resulted in roughly 10-20% repression in both cell types and the CMV-SV40 enhancer-promoter driven LacI-expressing plasmids resulted in ∼60% repression in both cell types. For each of the 3 plasmid combinations above, 100ng of pRL, 4:1 molar equivalents of repressible module plasmid:pRL, and 4:1 molar equivalents of LacI-expressing plasmid:pRL were co-transfected into 1 million K562 cells or PAC2 cells. 6 biological replicates were performed for each condition and 3 were exposed to 1mM IPTG and the remaining 3 were not exposed to IPTG. Error bars represent standard deviation of replicates (n=3). Points represent values for all 3 replicates in each condition. The dashed line shows the NF LacI IPTG- negative control.

5.4 Discussion

Zebrafish are a commonly used model organism for studying the spatio-temporal dynamics of cis-regulatory element activity and gene function. However, the flexible and widely used *lac* operator-repressor system has previously been untested in zebrafish. Here we demonstrate that the *lac* operator-repressor system functions in zebrafish cells, consistent with observed activity in other model eukaryotic systems.

For the development of a reporter system in whole organisms like zebrafish, it is critical to demonstrate non-tissue specific activity of an enhancer to provide robust output for single-cell reporter signal detection. The CMV enhancer is routinely

used in reporter assays to drive strong and constitutive gene expression. We provide quantitative evidence that the CMV-SV40 enhancer-promoter robustly increases luciferase gene expression over an SV40-only control plasmid in zebrafish fibroblasts. Furthermore, strong, non-tissue specific eGFP signal was observed in fertilized zebrafish eggs that persisted 72 hours post-fertilization after integration of a CMV enhancer controlled SV40-eGFP reporter construct. The robust levels of expression are critical in whole-organism studies where only a small number of cells may be expressing a reporter gene, and a high level of expression from a non-tissue specific enhancer-promoter such as CMV-SV40 may facilitate their detection.

Changes in expression of the reporter protein LacI are inversely related to changes in reporter expression. This response appears to provide a level of repression directly related to the LacI level rather than functioning as an on/off switch. This will allow for a more nuanced measure of *lac* regulatory control. Upon the addition of IPTG, luciferase signal was recovered to the levels of a non-functional LacI control, indicating that robust repression is completely reversible at low IPTG concentrations. The pronounced response to IPTG treatment, as well as minimal toxicity in a zebrafish cell line, suggest the *lac* operator-repressor system is a viable tool for use in whole zebrafish.

Lac operator-repressor systems can be used to control endogenous gene expression without interrupting native regulatory processes as *lacO* sites can be inserted in benign regions such as introns and UTRs. Transcriptional inhibition of RNA polymerase by steric hindrance can achieve repression without introducing artificial modifications to the locus and causing prolonged alterations in regulatory behavior. This is in contrast to other systems that achieve transcriptional control by tethering a protein domain with activating or silencing effects through chromatin

modifying or other endogenous mechanisms. Specificity of repression is also less of a concern compared to novel CRISPRi methods known for off-target effects [166]. As demonstrated by the REMOTE-control system, the *lac* system can also be used in conjunction with Tet-related systems to drive both activation and repression of a single loci, bringing additional flexibility to zebrafish studies [161]. This system also allows for time-controlled experiments, where a reporter gene is repressed only for a limited time window, making it a crucial tool for replicating the restriction of gene expression during development.

5.5 Methods

Plasmid design

CMV-SV40 enhancer-promoter luciferase plasmids were generated by restriction digestion to insert a CMV enhancer and a minimal SV40 promoter, or only a minimal SV40 promoter, upstream of a luciferase reporter molecule in the context of a pGL3 plasmid (Promega, E1751). Plasmids designed for whole zebrafish injection were generated by replacing the firefly luciferase reporter with an enhanced green fluorescent protein (eGFP) reporter, and adding flanking minimal Tol2 200 base pair 5' sequence and 150 base pair 3' sequence for integration into the genome [167]. The sequence of the CMV-SV40 promoter is as follows:

GGCATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTA
GTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGG
CCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGA
CGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGG
GTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCA

TATGCCAAGTCCGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCT
GGCATTATGCCCAGTACATGACCTTACGGGACTTTCCTACTTGGCAGTAC
ATCTACGTATTAGTCATCGCTATTACCATGGACTTGCATCTCAATTAGTC
AGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGC
CCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTAT
GCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAGTGAGG
AGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTC.

Lac operator-repressor system plasmids were created using the EMMA golden gate assembly method [168]. All plasmids assembled using EMMA have a backbone consisting of an ampicillin resistance gene and a high-copy-number ColE1/pMB1/pBR32-2/pUC origin of replication. Backbone elements are denoted by terminating dotted lines in all plasmid schematics (Figure 5.1, Figure 5.2, Figure 5.4, Figure 5.3). The EMMA toolkit was a gift from Yizhi Cai (Addgene kit # 1000000119) [168]. The *lacI* CDS and C-terminal NLS were cloned from the Addgene plasmid pKG215 and inserted into an EMMA entry vector to create an EMMA part. pKG215 was a gift from Iain Cheeseman (Addgene plasmid # 45110) [169]. A frameshift mutation was introduced by inserting an adenosine in the fourth codon of lacI to create a non-functional LacI (NFLacI) for use in control experiments. The LacI-expressing module contains a minimal SV40 promoter, the *lacI* gene, and a SV40 polyA tail, with or without the addition of an upstream CMV enhancer (Figure 5.2). The repressible reporter plasmid includes a CMV enhancer and a minimal SV40 promoter upstream of a firefly luciferase gene with symmetric *lac* operators inserted in its 5'UTR, terminated by a SV40 polyA tail. In order to maximize repression activity, six copies of the *lac* operators containing the sequence AATTGTGAGCGCTCACAATT were utilized in this study. This sequence is the "symmetric" *lac* operator that possesses tighter

binding with LacI than the canonical *lac* operator sequences [161].

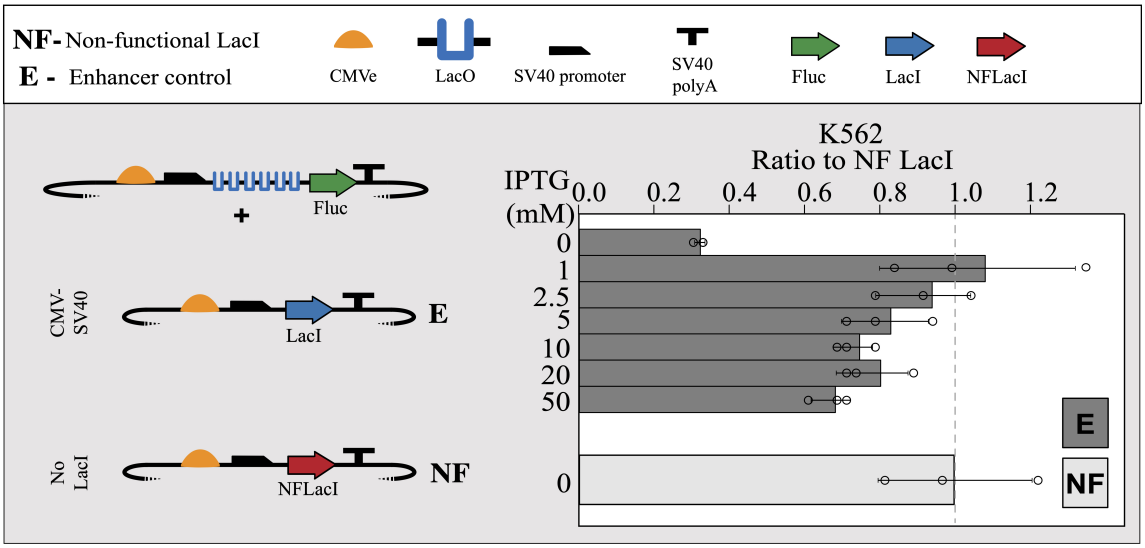


Figure 5.4: High levels of IPTG negatively impact the output of the *lac* operator-repressor system in K562 cells. Co-transfections of repressible reporter plasmids with equimolar amounts of CMV-SV40 enhancer-promoter enhancer driven LacI-expressing plasmids were exposed to increasing levels of IPTG. At 1mM IPTG, the output signal is restored to the level observed when NFLacI-expressing plasmids are co-transfected with repressible reporter plasmids, indicating that 1mM IPTG is sufficient to relieve LacI repression. At IPTG concentrations >1mM we observed an increasingly lower output of signal relative to NFLacI levels. Error bars represent standard deviation of replicates (n=3). Points represent values for all 3 replicates in each condition. The dashed line shows the NF LacI IPTG- negative control.

Cell culture

The zebrafish fibroblast cell line PAC2 was maintained as previously reported [170]. Cells were grown at 28 degrees Celsius in Leibovitz's L-15 +glutamine Medium (Invitrogen, 21083027) containing 15% heat inactivated fetal bovine serum (FBS; Sigma-Aldrich, F4135-500ML) and 1% antibiotic-antimycotic (Corning, MT30004CI) until confluent. Confluent cells were washed with 1x phosphate buffered saline (PBS; Invitrogen, 10010023) and detached from the plate with 0.05% Trypsin-EDTA for 5 minutes (Invitrogen, 25300054). Trypsin was quenched with FBS supplemented Leibovitz's L-15 Medium and detached cells were distributed into sterile flasks with

fresh media.

Electroporation and luciferase reporter assay

To assess the activity of CMV enhancer in zebrafish cell culture, 4000ng of firefly luciferase expressing plasmids either with or without the CMV enhancer were transfected into 2x106 PAC2 cells via electroporation (Figure 5.1A). 100ng of the renilla luciferase expressing plasmid (pRL-SV40 Promega, E2231) was included as a transfection control. Firefly/renilla luciferase signal was calculated as the mean of ratios of three technical replicates per biological replicate. Fold change was then calculated relative to the signal of the SV40 promoter-only containing plasmid. The mean of fold-changes is reported and error bars represent standard deviation. To test the functionality of our dual module *lac* repressor system in zebrafish cell culture, 2000ng repressible module and 2000ng of LacI-expressing plasmid were co-transfected into 2x106 PAC2 cells by electroporation. 400ng of pRL-SV40 was included as a transfection control (Figure 5.2). All transfections were completed using 2mM cuvettes (Bulldog Bio, 12358-346) and electroporated using a NEPA21 Electroporator (Nepagene). Cells were harvested from culture and resuspended in 90uL of Opti-MEM Reduced Serum Medium (ThermoFisher, 31985062) per 1x106 cells. Mastermixes of cells and DNA were prepared according to scale of conditions, and distributed into cuvettes (100uL/cuvette, 10uL of DNA and 90uL of cells). Poring pulse for PAC2 cells was set to the following: 200V, Length 5ms, Interval 50ms, Number of pulses 2, D rate% 10, Polarity +. Poring pulse for K562 was set to the following: 275V, Length 5ms, Interval 50 ms, Number of pulses 1, D rate 10%, Polarity +. For both cell types, the transfer pulse conditions were set to the following: 20V, Pulse length 50ms, Pulse interval 50ms, number of pulses 5, D rate 40%, Polarity +/-. Immedi-

ately following electroporation, each cuvette was recovered in 900L of appropriate media and distributed into a well on a 24-well culture plate. For the transfection in Figure 2, PAC2 cells were recovered in 6-well plates. Each condition had a total of 3 biological replicates. For experiments including LacI, IPTG was treated as a separate condition and added to 1mM final concentration, unless otherwise specified, at 1 hour and 24 hours post-transfection (Figure 5.4). Luciferase results were collected 48 hours post-transfection on a GloMax-Multi+ Detection System (Promega, E7081) using the Promega Dual-Glo Luciferase Assay System (Promega, E2940).

Zebrafish microinjections

Microinjections were carried out using the Tol2 transposon system as previously described [167, 171]. Zebrafish embryos were injected from a master mixture of 2uL of 125ng/uL assay plasmid, 2uL 175ng/uL Tol2 mRNA, 1uL phenol red dye, and 5uL sterile water within 30 minutes of fertilization. All embryos were maintained in Holt buffer and fluorescent activity assessed at 24, 48, and 72 hours.

5.6 Notes & Acknowledgments

This chapter was previously published to *bioRxiv* in February, 2020 [172]. The work presented here represents a group effort. I performed experiments and analysis in whole zebrafish. Torrin L McDonald, Gregory A Farnum, and Monica J Holmes carried out experiments in Pac2 cells. Torrin L McDonald, Gregory A Farnum, Monica J Holmes, Melissa L Drexel, and Jessica A Switzenberg assisted in plasmid conception and creation.

CHAPTER VI

Novel Inversion Assays for the Study of Negative Regulatory Elements in Whole Zebrafish

6.1 Abstract

Transcriptional regulation by non-coding cis-regulatory elements is a key process driving tissue-specific gene expression during development. While the identification and characterization of promoters and enhancers, also known as positive regulatory elements, has led to a better understanding of mechanisms of gene regulation and variation associated with human disease, negative regulatory elements such as silencers and enhancer blockers are comparatively understudied. This is because current plasmid-based reporter assays used to identify and characterize cis-regulatory elements are optimized for positive regulatory elements, but provide only a reduced or negative reporter signal output for negative regulatory elements. Measuring a loss of signal makes these assays susceptible to high rates of false positives and thus impracticable for characterizing negative regulatory activity in whole animals. Here we describe the first reporter assay that produces a positive reporter output in response to negative regulatory element activity. This assay can be used to identify negative regulatory elements in cell lines and define the spatio-temporal activity of negative regulatory elements in whole organisms. We demonstrate that this assay can be used to assess negative regulatory activity in cell lines and potentially function *in vivo*.

This advancement will allow for more reliable identification of negative regulatory elements, supporting their identification genome-wide in a manner similar to what has already been achieved for PREs, dramatically expanding our understanding of genome-wide gene regulation.

6.2 Introduction

Over the last 30 years, the discovery and characterization of cis-regulatory elements has expanded our understanding of gene function and disease mechanisms. It is now widely accepted that variants falling into non-coding regulatory sequences have the ability to alter gene expression in ways that disrupt gene function, but are less likely to be embryonic lethal compared to genic variants. This is supported by genome-wide association studies (GWAS) that have found the majority of disease-associated variation falls into non-coding sequence [3, 82]. In addition, regulatory sequences are known to control tissue and time-point specific gene expression during development. Efforts supporting the discovery and characterization of cis-regulatory elements in the genome are necessary for the full delineation of human development and disease-related risk loci.

The majority of characterized regulatory elements are associated with positive target gene expression. These include promoters and enhancers, also known as positive regulatory elements (PREs), and have been the focus of many studies and technological advances. Traditionally PREs have been studied using plasmid-based reporter assays, where a putative PRE is inserted upstream of a reporter gene, such as GFP [29]. If the putative PRE has activity consistent with driving or increasing gene expression, a positive GFP fluorescent signal can be detected. These assays

have been utilized successfully to identify many PREs as well as characterize their spatio-temporal activity *in vivo*, as seen in studies of human limb malformations [83]. Large scale efforts to characterize enhancers have been untaken using reporter assays, as shown in the VISTA Enhancer Browser [31]. More recently, high-throughput advancements of reporter assays, like massively parallel reporter assays (MPRAs) and STARR-seq, have allowed for the identification of PREs at unprecedented rates [33, 32, 173]. These assays have permitted the classification of histone modifications associated with PREs, further facilitating their identification by providing parameters to prioritize likely candidates [174, 175].

Despite the abundance of tools available for regulatory element testing, very few of these elements have been associated with a decrease in target gene expression [176]. These negative regulatory elements (NREs) include enhancer blockers, which disrupt enhancer-promoter communication in a position-specific manner, and silencers, that potentially decrease gene expression through a variety of mechanisms, including premature termination of transcriptional elongation (Figure 6.1) [177, 178]. There are a predicted >1.7 million silencer elements alone within the human genome, however very few of these have been extensively characterized or functionally validated [179, 180]. In fact, there are not yet enough validated NREs to confidently predict additional NREs via histone modifications [181, 179, 180, 182, 183]. Multiple variants fall within NREs and have been associated with various human diseases, including muscular dystrophy, asthma, blood-pressure control, and Beckwith-Wildemann syndrome. The lack of characterized NREs limits our understanding of genomic mechanisms leading to human disease [184, 185, 186]. Extensive efforts to identify and characterize NREs need to be implemented to advance our understanding of gene regulation and disease mechanisms as a whole.

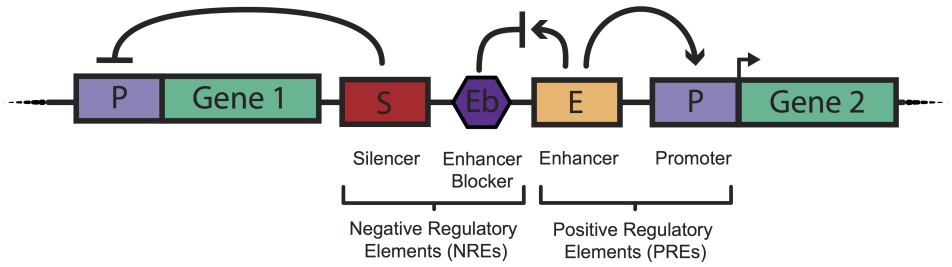


Figure 6.1: Functions of four types of regulatory elements. Positive regulatory elements include promoters, which initiate transcription, and enhancers, which increase gene expression. Negative regulatory elements include silencers, which decrease gene expression, and enhancer blockers, which interfere with enhancer activity on their target promoters. S, silencer; Eb, enhancer blocker; E, enhancer; P, promoter. Adapted from Vandermeer et al. [83].

Though recent efforts have been made to identify NREs in the genome using high-throughput screens and prediction models, additional tools for validating NREs are still required [179, 180, 187]. Low-throughput reporter assays, similar to those used for PREs, have been applied to NRE characterization [34, 188]. These assays are structured similarly to classic enhancer assays with the addition of an insertion site for either a silencer upstream of the enhancer, or an enhancer blocker between the enhancer and promoter. The resulting NRE activity is interpreted by a reduction or complete ablation of signal from an initially moderate or high positive baseline signal. This reliance on a reduction or loss of signal in these reporter assays makes them vulnerable to false positives through mutated plasmids or failed transfections, and false negatives though the masked signal of weak silencers or incompatible cell lines or molecular machienery, and consequently has prevented them from being adopted for widespread use. Evidence suggests that NREs may have tissue-specific activity, however this activity remains difficult to assess when these assays are utilized *in vivo*, as a small region of reduced fluorescence must be distinguished from the background signal of a fully fluorescent animal [181, 189, 187]. Recent efforts have been made to characterize the spatio-temporal activity of silencers in whole Drosphila using

a classical silencer assay. The drawback is these assays are limited to silencing expression driven by a cell-type specific enhancer, and are not easily scalable for comprehensive studies of NREs within an organism [181].

In order to address the current limitations of NRE assays, we have developed a panel of reporter assays that output a positive reporter signal in the presence of active NREs. Two versions have been developed, using either CRISPR or the *lac* operator-repressor system as a functional intermediate unit to generate double-negative assays. The typical reduction of expression, stimulated by the presence of a candidate repressive element, is linked to the increased expression of a reporter gene through the use of these intermediates. This essentially inverts the NRE signal, turning negative regulation into positive expression. This advancement allows for characterization and validation of the spatio-temporal activity of NREs, both in cells and in whole animals. The CRISPR-based inversion assay is utilized here to validate the activity of NREs in cells and demonstrate the tissue-specific activity of a known silencer element in a transgenic zebrafish model organism.

6.3 Methods

dCas9 assay construction

The dCas9 plasmid was generated using dCas9 from the pHR-SFFV-KRAB-dCas9-P2A-mCherry which was a gift from Jonathan Weissman (Addgene plasmid # 60954) [166]. The self-cleaving ribosomes were synthesized as separate gBlocksand cloned into pENTR1A (Thermo Fisher). Tol2 sites were taken from the Tol2kit plasmid #396 [190], and SV40 promoter from the pGL3.promoter commercial plasmid (Promega).

Our deactivated dCas9 (ddCas9) was generated using single site mutagenesis to insert a thiamine directly after the ATG start codon, causing a frameshift mutation and generating a stop codon.

lac operator-repressor assay construction

The *lac* operator-repressor plasmids were constructed using EMMA Mammalian Modular Assembly (EMMA) golden gate assembly. The EMMA toolkit was a gift from Yizhi Cai (Addgene kit # 1000000119) [168]. The *lac* operator-repressor system components were cloned from the Addgene plasmid pKG215. pKG215 was a gift from Iain Cheeseman (Addgene plasmid 45110) [169]. The SV40 promoter and the CMV enhancer sequence is as previously published [172]. The version of LIRA presented here contains six copies of the *lacO* sequence: AATTGTGAGCGCTCACAATT.

Cell Culture

The human myelogenous leukemia cell line K562 was maintained as previously reported by ATCC and briefly described here. Cells were grown at 37 degrees Celsius and 5% CO2 in RPMI-1640+L-glutamate media (Sigma-Aldrich, 11875093) containing 10% heat inactivated fetal bovine serum (Sigma-Aldrich, F4135) and 1% antibiotic-antimycotic (Corning, MT300004CI)(RPMI 1640 complete media).

The zebrafish fibroblast cell line PAC2 was maintained as previously reported [170]. Briefly, cells were grown at 28 degrees Celsius in Leibovitz's L-15 +glutamine media (Invitrogen, 21083027) containing 15% heat inactivated fetal bovine serum (Sigma-Aldrich, F4135) and 1% antibiotic-antimycotic (Corning, MT300004CI).

Cell Transfections

All transfections into K562 cells were completed using 2mM cuvettes (Bulldog Bio, 12358-346) and electroporated by a NEPA21 Electroporator (Nepagene). Cells were harvested from culture and resuspended in 90uL of Opti-MEM Reduced Serum Medium (ThermoFisher, 31985062) per 1x106 cells. Cells and DNA were combined in a mastermix before distribution into cuvettes (100uL/cuvette: 10uL of DNA and 90uL of cells). Nepagene poring pulse was set to 275V, Length 5ms, Interval 50 ms, Number of pulses 1, D rate 10%, Polarity +. Nepagene transfer pulse was set to 20V, Pulse length 50ms, Pulse interval 50ms, number of pulses 5, D rate 40%, Polarity +/-. Immediately following electroporation cells were recovered in 900uL RPMI 1640 complete media. Cells were then grown in 24-well culture plates and RNA was harvested after 48hrs. All conditions had 2 biological replicates.

qPCR

RNA extraction was completed using the RNeasy mini kit (Qigen), and cDNA reverse transcription was carried out using the High Capacity RNA-to-DNA kit with oligo dT (Applied Biosystems). Quantitative PCR (qPCR) was carried out using Power SYBR Green master mix (Applied Biosystems, 4369659).

Zebrafish microinjection

Microinjections were carried out as previously described [167, 171]. Zebrafish embryos were injected from a master mixture of 2uL of 125ng/uL assay plasmid, 2uL 175ng/uL Tol2 mRNA, 1uL phenol red dye, and 5uL sterile water within 30 minutes of fertilization. All embryos were raised in 1x Holt buffer. Fluorescent activity was assessed at 24 and 48 hours post fertilization (hpf).

6.4 Results

dCas9 positive NRE assay

Here we introduce the first assay to characterize NRE activity in whole animals using a positive readout for repressive activity. The creation of this inversion assay takes advantage of the newly developed catalytically dead Cas9 (dCas9) protein as a targeted transcriptional silencer of the reporter molecule enhanced GFP (eGFP) [188]. The assay functions by placing a sgRNA recognition sequence in the 5' UTR of the eGFP gene, allowing for the dCas9-sgRNA complex to bind. As the dCas9 is catalytically dead, it will not cut when bound to this recognition sequence, and instead blocks transcription elongation by steric hindrance (Figure 6.2). However, when an active NRE is inserted upstream of the sequence encoding the sgRNA, it reduces the available sgRNA that can complex with dCas9, allowing eGFP to be expressed. Through this method, NREs are able to produce a positive reporter signal consistent with the cell-type specificity of the NRE being tested. While transcription of sgRNA is typically driven by a PolIII promoter in other assays, we chose to modify ours to be compatible with PolII transcriptional control to maintain a regulatory environment known to be compatible with the majority of regulatory element activity in eukaryotes. A polyA sequence was added to the 3' end of the sgRNA, which is removed post-transcriptionally through the inclusion of flanking self-cleaving ribozymes along with the excess 5' sequence [191]. These ribozymes are capable of cleaving themselves to free the sgRNA post-transcriptionally and allow it to complex properly with the dCas9 protein.

We designed four versions of this plasmid, one to detect the activity of each of

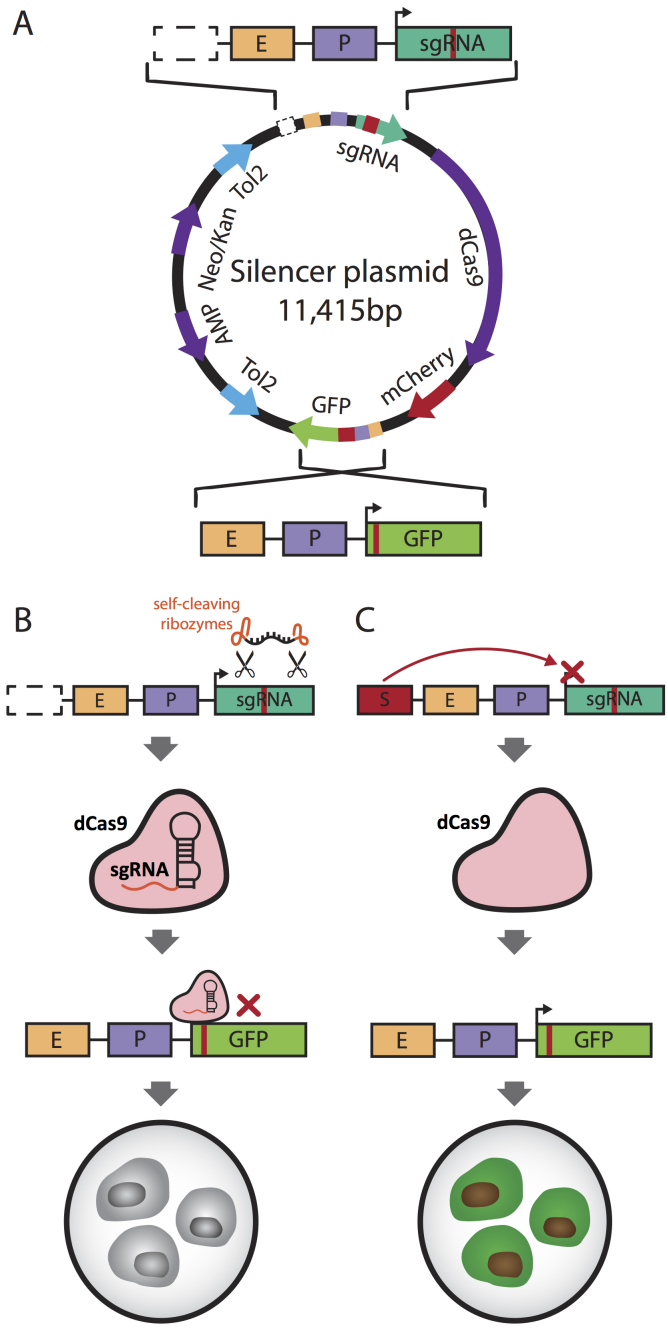


Figure 6.2: dCas9-based positive assay for negative regulatory elements. These assays use a catalytically dead Cas9 (dCas9) as a targeted transcriptional silencer of a ubiquitously expressed reporter molecule. A putative silencer or enhancer blocker is inserted (dashed box) upstream of the regulatory machinery for an sgRNA which targets the 5' end of the reporter molecule (eGFP). eGFP is only expressed when an active negative regulatory element is present, repressing sgRNA expression which prevents dCas9 targeting of eGFP. Alternatively, if the sequence inserted is not an active negative regulatory element, sgRNA is expressed and complexes with the constitutively expressed dCas9, targeting it to eGFP. The dCas9-sgRNA acts as an intermediate cassette to invert the negative element's signal, yielding a positive change in reporter expression. E, enhancer; P, promoter.

the four types of known regulatory elements (Figure 6.3). For NREs, the silencer assay contains gateway sites that allow easy insertion of a putative silencer element upstream of the enhancer and promoter driving sgRNA expression [192]. Similarly, our enhancer blocker assay contains gateway sites between the enhancer and promoter driving sgRNA expression. The PRE assays are mechanistically equivalent to classical promoter and enhancer assays, and use the same backbone as the NRE assays. This is to maintain size and content consistency between plasmids to avoid bias when comparing putative regulatory elements across all four assays. Putative PREs can be inserted upstream of the minimal eGFP promoter (enhancer assay), or in place of the minimal eGFP promoter (promoter assay). To remove any transcriptional repression of eGFP from dCas9 in these plasmids, mutagenesis was utilized to generate a frameshift mutation leading to a truncated "deactivated" dCas9 protein (ddCas9).

dCas9 inversion assays correctly assess activity of known silencers and enhancer blockers

Plasmids of all four dCas9 assay types were constructed to contain five distinct test regions: the HS2 enhancer, a conserved enhancer blocker, a non-conserved enhancer blocker, a conserved silencer, and a non-conserved silencer [34]. These plasmids were separately transfected along with two negative controls for the enhancer/promoter (E/P) assay and silencer/enhancer blocker (S/EB) assay into K562 human myelogenous leukemia cells. By quantifying the resulting eGFP expression by qPCR, we found regulatory elements to show the highest levels of expression in plasmids matching their expected regulatory type (Figure 6.4). This suggests that the 4 plasmid model is able to accurately delineate the regulatory function of tested cis-regulatory

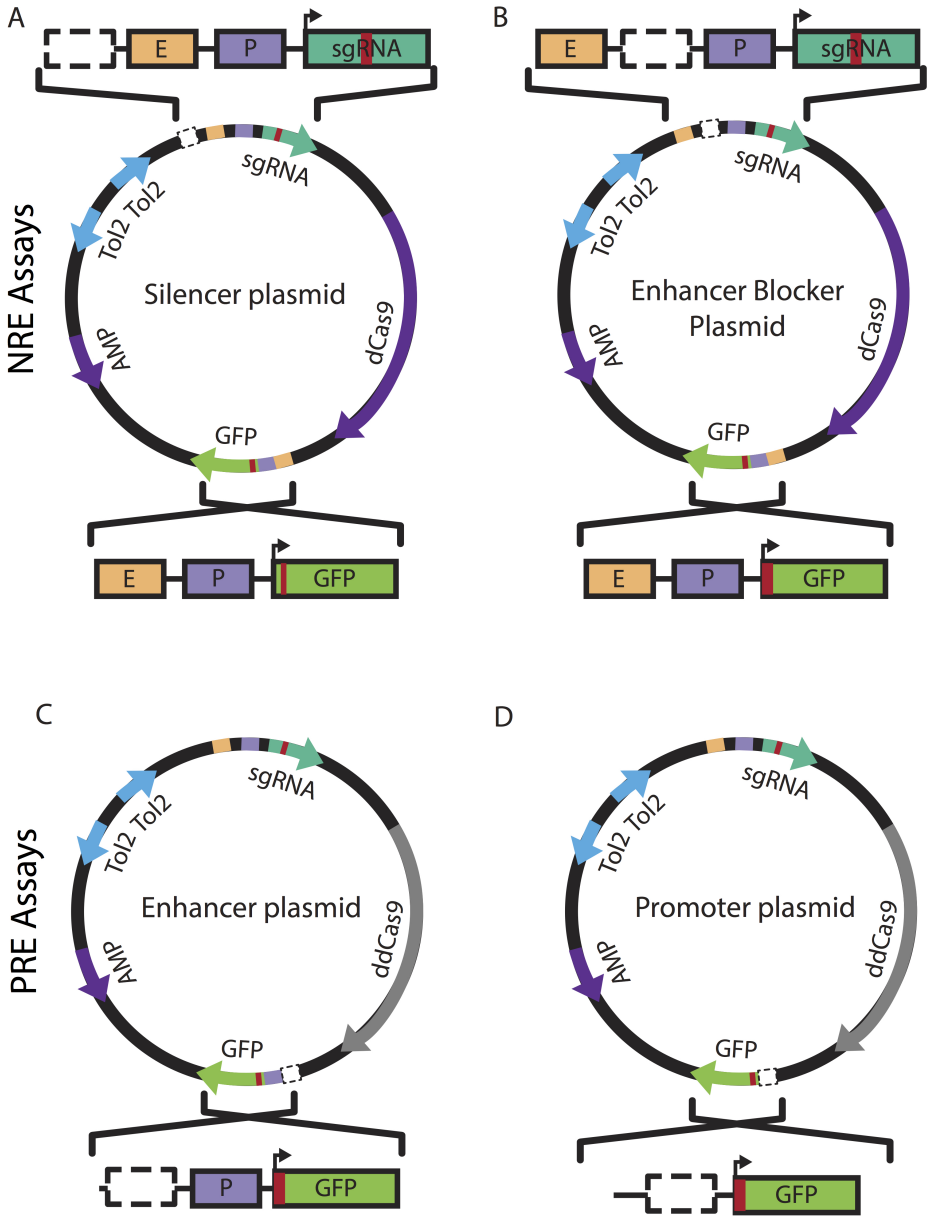


Figure 6.3: dCas9 inverted reporter assays for four regulatory element types. E, enhancer; P, promoter; dCas9, catalytically dead Cas9; ddCas9, deactivated dCas9; red line, sgRNA barcode recognition sequence; dashed box, gateway insertion site.

elements. While the majority of the NREs tested primarily showed activity in their matched assay type, the conserved silencer element demonstrated activity in both the silencer and the enhancer blocker assays, consistent with previous characterizations of silencers as orientation-independent [193].

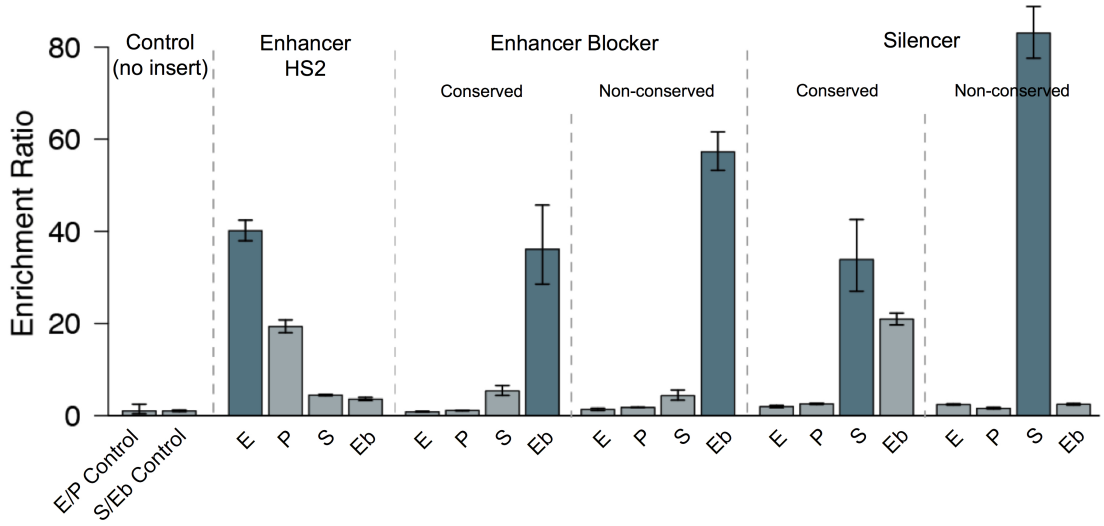


Figure 6.4: Control positive and negative elements demonstrate expected activities in dCas9-inverted reporter assay panel. Five previously-tested control regulatory elements were tested in each of our dCas9 assays. eGFP transcription levels were assessed by qPCR 48 hours post transfection into K562 cells. Control elements tested include the HS2 enhancer, a conserved enhancer blocker, a non-conserved enhancer blocker, a conserved silencer, and a non-conserved silencer [34]. Dark bars represent the highest enrichment ratio for each regulatory element. All elements tested generated the highest reporter levels when placed in their corresponding assay position, i.e. enhancer blockers show strongest activity when placed in the enhancer blocker position between enhancer and promoter. All values shown are proportional to background reads from empty enhancer and promoter (E/P) or silencer and enhancer blocker (S/Eb) controls. X-axis labels represent the type of dCas9 assay used. Labels above the bars indicate the tested element. Error bars represent standard deviation of 2 biological replicates. E, enhancer assay; P, promoter assay; S, silencer assay; Eb, enhancer blocker assay.

in vivo analysis of NRE in dCas9 inversion assay in zebrafish shows highly mosaic expression

One of the largest advancements of these assays for NREs is the ability to utilize them in model organisms to examine *in vivo* spatio-temporal expression patterns of these elements. Zebrafish was chosen as a model system as they have historically been used to assess the activity of enhancers via *in vivo* enhancer assays [29]. Similar to PREs, functional conservation of NREs has been shown previously across species, suggesting expression patterns produced from these assays will likely be conserved to humans [194, 195, 196]. Zebrafish represent a convenient model organism for studies

yielding a tissue-specific fluorescent output due to their high level of conservation to human genic regions, large numbers of offspring per mating, rapid development, and transparent bodies in early development [167].

Plasmids containing the known non-conserved silencer previously tested by qPCR were microinjected into 302 one- or two-celled zebrafish embryos (86 uninjected controls), and eGFP expression patterns were qualitatively analyzed at 24 and 48 hours post fertilization (hpf) [167, 34, 171]. 69.8% (37/53) of our injected embryos were eGFP positive and no eGFP activity was seen in the uninjected controls (0/8) at 24 hpf. Highly mosaic fluorescent activity across all eGFP positive zebrafish embryos was observed (Figure 6.5). This mosaicism is likely due to the >12kb plasmid size, as integration efficiency of the Tol2 recombination system is known to decrease with plasmid size [197]. An mCherry gene was included on the backbone of the plasmid assay as a transfection control. Some cells in the zebrafish embryos expressed both mCherry and eGFP, signifying a tissue type in which the tested silencer is active. However, some cells in these fish were observed to express mCherry but not eGFP, representing tissue types where the tested silencer is inactive. These results strongly support tissue-specificity of this silencer element, and broadly reinforce previous evidence of tissue-specific NREs [181, 189, 187].

The Lac Inverted Reported Assay

The highly mosaic fluoresce seen following integration of the dCas9 silencer assay in zebrafish embryos is consistent with low transfection efficiency of the plasmid and size limitations of the Tol2 system. To address this limitation, an alternate repression system was utilized to serve as the inversion mechanism driving our reporter assay. The dCas9 assay was substituted with a *lac* operator-repressor system model, which

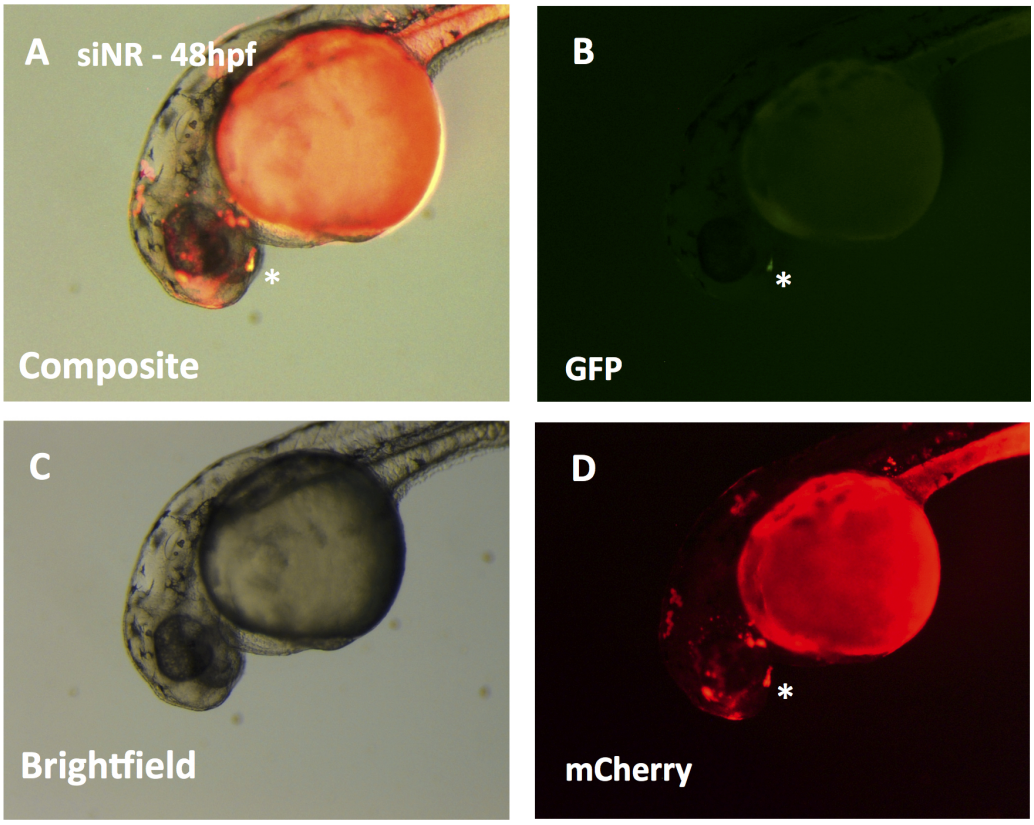


Figure 6.5: Fluorescent signal in zebrafish 48 hours post-fertilization(hfp) reflects tissue-specific silencer activity. A. Composite image shows a * where tissue specific active silencer activity can be seen (yellow). B. Following microinjection of a non-conserved known silencer element from Petrykowska et al. silencer activity was observed, signified by eGFP expression [34]. C. Brightfield zebrafish. D. Tissues in which our assay has been integrated into the genome are signified by mCherry expression.

has known reporter signal inversion capabilities, and reduces the genomic size of our plasmid-based assay compared to the dCas9 model [161].

The *lac* operator-repressor system functions through the expression of a Lac repressor protein (LacI) that specifically binds a *lac* operator sequence (*lacO*) [152]. When *lacO* sites are placed upstream of a reporter molecule, such as eGFP, and LacI protein is present, eGFP transcription is inhibited via steric hindrance through the LacI-*lacO* site binding. Our assay uses a classical NRE assay to drive LacI expression, where an active NRE is expected to reduce the levels of LacI protein, reducing LacI-*lacO* site binding, therefore allowing eGFP expression (Figure 6.6).

We previously established the functionality of *lac* operator-system in the zebrafish fibroblast cell line PAC2, supporting the potential function of this system in whole fish [172]. We also included cHS4 insulators between the expression cassettes in the assay to insulate potential secondary enhancer activity [198]. These insulators are well characterized CTCF-based dual function insulator-enhancer blockers shown to have strong activity across multiple cell types. This new NRE assay is termed the Lac Inverted Reporter Assay (LIRA)(Figure 6.7).

LIRA still requires validation to confirm that this method functions as expected in whole zebrafish. The first planned assays for microinjection with Tol2 include a positive control plasmid expressing LacI, and a negative control plasmid with no functional LacI expression (Figure 6.8). In the positive control plasmid, no NRE will be inserted upstream of LacI, allowing positive LacI expression and resulting no eGFP expression. This positive control will demonstrate the function of the *lac* operator-repressor system in whole zebrafish by revealing any tissue-specific eGFP silencing by LacI. It is expected that all cell types in which the *lac* operon-repressor system functions will be eGFP negative. In the negative control plasmid, no LacI expression or eGFP expression is expected as it lacks an enhancer or putative NRE upstream of LacI. In addition, the negative control will reveal any tissue-types in which the enhancer element driving LacI is non-functional. Any tissue-types where the enhancer is unable to drive LacI expression will be able to express eGFP. Addition of the allosteric inhibitor Isopropyl β-d-1-thiogalactopyranoside (IPTG) may also be used with the negative control fish to release LacI binding to *lacO* sites, functionally reversing the repression driven by the *lac* operator-repressor system to reestablish eGFP expression. The successful reversal of eGFP repression will further support the functional role of the *lac* operator-repressor system in the LIRA assay as well as

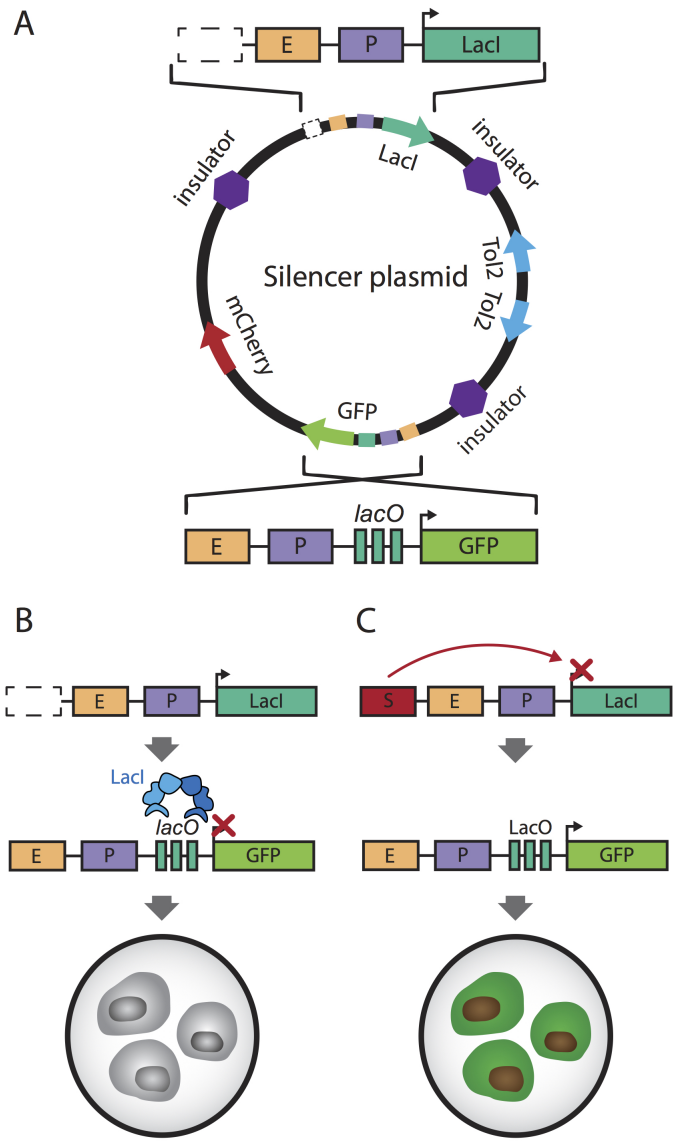


Figure 6.6: Lac Inverted Reporter Assay (LIRA) for the assessment of NRE activity. Our LIRA inversion assay produces positive eGFP signals from active negative regulatory elements using the *lac* operator-repressor system. In the silencer assay, when no silencer is inserted as part of the LacI cassette, LacI protein is expressed. This LacI protein specifically binds to the *lacO* recognition sequences in the 5' UTR of eGFP, sterically hindering eGFP expression. Alternatively, when a silencer is inserted, LacI expression is reduced and LacI protein is unavailable to bind to its *lacO* sites, allowing transcription of eGFP. The *lac* operator-repressor system acts as an intermediate to invert the silencer signal such that strong repressive activity produces strong eGFP expression. LacI plasmid E, enhancer; P, promoter; dashed box, gateway insertion site.

in whole zebrafish. The negative control will determine whether components of this assay, other than the presence of LacI, are responsible for driving aberrant eGFP

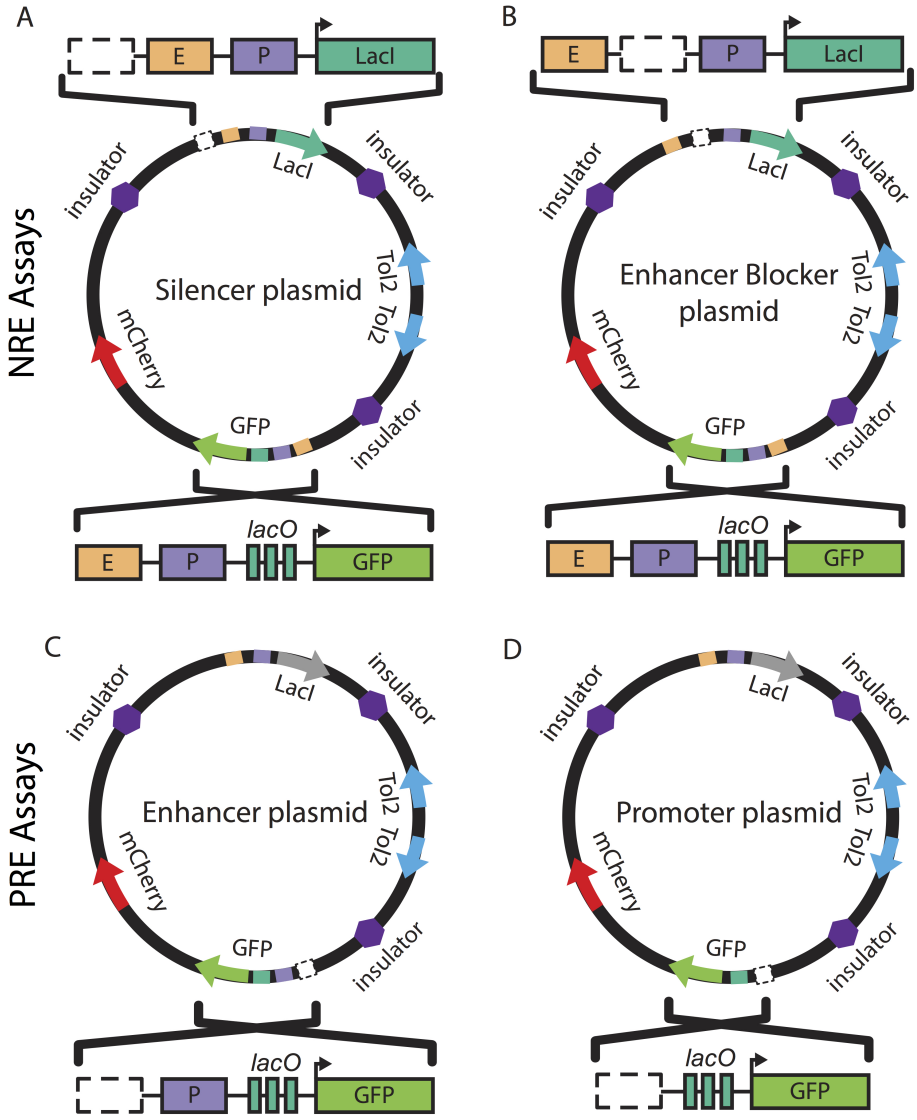


Figure 6.7: LIRA assays for four regulatory element types to assess the potential regulatory activity type of any putative regulatory element. Gray LacI represents a non-functional LacI gene containing an early frameshift mutation. E, enhancer; P, promoter; dashed box, gateway insertion site.

repression.

Following the observation of expected eGFP patterns within zebrafish controls, microinjections of NRE plasmids containing known NREs will proceed. We plan to start by examining conserved NREs identified by the classical NRE assays from Petrykowska et al. [34]. As these NREs were classified as active in the K562 ery-

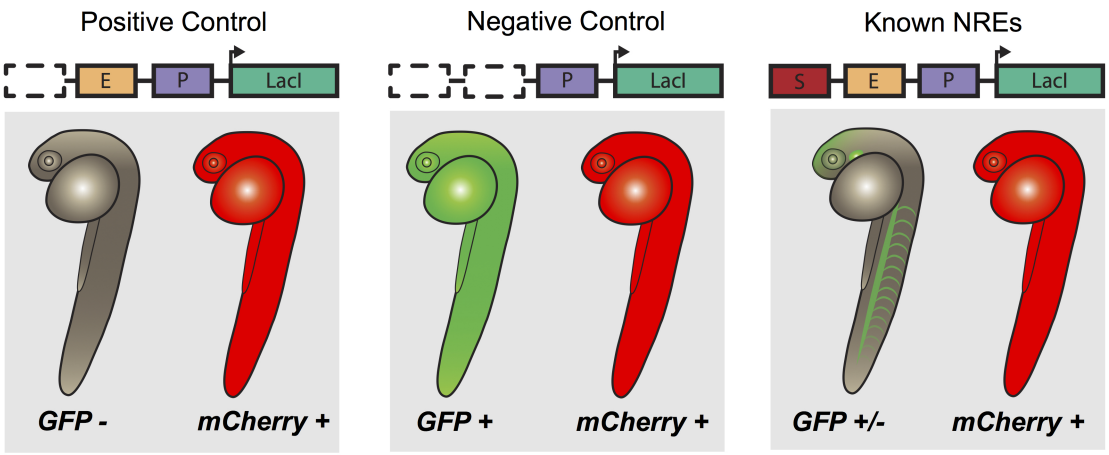


Figure 6.8: LIRA control plasmids and expected outcomes. Control plasmids will be required to test the function and feasibility of the LIRA assay system. A positive control, containing no NRE, is expected to yield no eGFP signal. A negative control, containing no NRE or enhancer driving LacI expression, is expected to yield ubiquitous eGFP expression. mCherry functions as a transfection control and is expected to be actively expressed in all cells which have successfully integrated the LIRA construct. E, enhancer; P, promoter; dashed box, gateway insertion site.

throleukemia cell line, we hypothesize that these control regions will be active in zebrafish blood cells, but may not be blood cell specific. The LIRA plasmid also includes constitutively active mCherry cassettes, allowing visualization of all cells in which the assays have successfully integrated into the genome to determine the level of mosaic integration in whole zebrafish.

6.5 Discussion

Fully elucidating the spatio-temporal activity of NREs is a key step in understanding regulatory processes and disease mechanisms. Here, we describe novel assays to characterize the spatio-temporal activity of NREs in cell culture and in a transgenic zebrafish model system. These assays use either dCas9 or the *lac* operator-repressor system to invert NRE negative signal into positive reporter expression, overcoming limitations of past NRE-based assays, and making these the first assays of their kind.

The dCas9-based assays produced expected activity when tested using previously annotated regulatory elements, including an enhancer, silencers, and enhancer blockers. This result provides confidence in our assays' ability to define regulatory elements by their biological function. In addition, we demonstrated tissue-specific silencer activity of a known silencer by using our dCas9 silencer assay in a zebrafish model organism. Going forward, the LIRA assays are expected to provide a more complete view of tissue-specific NRE activity by reducing transgenic mosaicism within the zebrafish models. We have developed novel assays capable of identifying the type of regulatory activity of a putative regulatory element, as well as characterizing its tissue-specific activity *in vivo*. These assays have the potential to rapidly expand the pace of validation and characterization of NREs, enhancing our knowledge of gene regulation.

The application of our assays to NRE testing across cell-types and organisms can resolve key questions regarding the behavior of NREs. Though some elements have been found to have dual silencer/enhancer blocker activity depending on cell type, it remains unknown how widespread this bifunctional activity is [199, 181, 200, 201]. By utilizing both our silencer and enhancer assays in zebrafish microinjection for the same putative bifunctional element, any differences in regulatory activity between tissue types can be visualized. Additionally, some NREs have shown orientation-dependent behavior, where they only function in certain orientations relative to a promoter or enhancer [138, 202, 203, 176, 193]. By placing NREs in forward or reverse orientation within our assays, this orientation-dependency of NREs can be fully characterized. Due to the rapid nature of the method used for plasmid assembly, integrating diverse promoters and enhancers into our plasmids to test for promoter or enhancer specific NRE activity is also easily attainable [204, 205, 206].

Our reporter assay is subject to the same limitations as other plasmid-based assays. Plasmids are commonly used due to the ease of generation and transfection, however they require testing outside native chromatin contexts and may not entirely reflect the native behavior of tested elements. Silencers have been known to function through a variety of mechanisms, and certain types of silencing activity may not be detectable in our assay, such as those thought to inhibit splicing [207, 208]. This will likely have less of an affect on enhancer blockers, some of which are thought to function through DNA structural changes [209, 210]. Given that our assay is constructed *in vitro*, it does not recapitulate endogenous patterns of DNA methylation or histone modifications, both of which may be required for certain NRE repressor-protein interactions [180]. Going forward it will be critical to validate that our assays function in whole zebrafish in addition to PAC2 cells, in order to allow us to characterize any cell-type specific expression driven by negative regulatory elements. It is yet to be determined if the enhancer used in our assay drives fully ubiquitous eGFP expression in whole zebrafish. We expect to observe false positive GFP expression in tissue-types without enhancer activity, which would functionally appear as a silenced or blocked enhancer area. Therefore it is imperative to examine our negative control using transgenic zebrafish to reveal any tissues where the enhancer may not be functional and actively expressing eGFP in the absence of a NRE. Additionally, our positive control must be analyzed in fish to ensure the eGFP background signal is low enough to not disrupt NRE assessment. Further reduction in NRE signal can be achieved in our LIRA assays through the inclusion of additional *lacO* sites in the 5' UTR of eGFP if this is an issue [161]. In addition, the use of a zebrafish model organism has its own limitations, given that the regulatory elements we test will likely need to be conserved from human to zebrafish in order to give us confidence

in the conserved function.

These assays can be adapted to identify silencers in an unbiased fashion genome-wide by extending another version of enhancer reporter assays, the enhancer trap. This method takes advantage of the random integration of Tol2 to insert an empty enhancer assay construct so that nearby enhancers are able to trigger reporter expression at the integration site [211, 212, 213, 214]. By adapting this methodology for use in our assays, we can identify novel silencers in the genome in an unbiased fashion. Through integration of these results with Hi-C measures of chromatin organization, we can utilize identified NREs to generate maps of regulatory networks within topologically associated domains (TADs) in the genome. This will provide a previously unattainable comprehensive overview of gene regulation [215]. By taking advantage of the barcoded system within the dCas9 assay, we can generate a high-throughput version of our assays. Genomic fragment libraries can be inserted into the gateway sites, and transfected at a low copy number into cells allowing for fluorescence-activated cell sorting. Isolated eGFP positive cells can be sequenced to discover novel NREs driving eGFP expression in an unbiased fashion.

In this paper we introduced two novel methodologies which allow NREs to be functionally characterized for spatio-temporal activity simultaneously across a whole organisms for the first time. Reporter assays optimized for the study of NREs will allow for more efficient validation and characterization of these elements in a time where their discovery is rapidly increasing. Better NRE representation will provide a more comprehensive understanding of gene regulation as a whole, as well as uncover novel regulatory sequences in the genome potentially contributing to human disease.

6.6 Notes & Acknowledgments

The work presented here represents a group effort. I performed experiments and analysis in whole zebrafish. Jessica A Switzenberg carried out the qPCR analysis. Torrin L McDonald, Gregory A Farnum, Monica J Holmes, Melissa L Drexel, and Jessica A Switzenberg contributed to plasmid conception and creation.

CHAPTER VII

Conclusions and Future Directions

Gene regulation is a key eukaryotic genomic process that allows the precise transcription of genes in a cell-type specific manner. Genomic regulatory regions are enriched for variants associated with a number of human diseases, including multiple cancers and neurological disorders. Despite their importance to the interpretation of the human genome and prediction of human disease, gene regulatory regions throughout the non-coding genome remain vastly understudied.

The work presented in this dissertation advances non-coding genomic regulatory sequence studies by introducing novel computational tools to examine variation and methylation consequences within transcription factor binding sites, and the development of novel experimental assays for regulatory element characterization. These tools are expected to contribute to the study of gene regulation by providing accessible predictions and spatio-temporal characterizations of functional regulatory sequence. The resulting datasets will help advance research and experimental validation of these understudied sequences.

7.1 Improving Predictions of Non-coding Variant Function

The work presented in Chapters 2-4 provide background on the current field of non-coding variation function prediction, as well as introduce novel tools advancing these predictions for transcription factor binding sites.

There remains a large pool of unvalidated putative disease causing non-coding variants identified by genome-wide association studies. This is due to experimentally rigorous methods used to validate putatively functional non-coding variation, such as fine mapping, that is impractical to implement for large numbers of variants. To address this drawback, multiple computational tools have been developed to assist in the prioritization of non-coding variants for further study. These tools employ genome-wide functional annotations and measures of conservation to generate predictions either heuristically, or using a machine learning based approach. In Chapter 2, I examined four of these tools (DeepSEA, RegulomeDB, CADD, and FATHMM-MKL) and revealed disparities in agreement between their functional calls [24, 27, 57, 58]. Previous implementations of non-coding annotation methods found they perform better in tandem, suggesting that combining annotation tools may lead to improved predictions of functional non-coding variation [109]. This result, combined with the finding that fewer than 40% of GWAS analyses in 2015 utilized a non-coding variation annotation method, suggest that improved computational predictions through the addition of functional annotations may further assist in the prioritization of non-coding variants for experimental validation.

To aid in this goal, I developed the computational method SEMpl, which uses functional annotations to generate predictions of variation falling within transcription factor binding sites genome-wide, and is described in Chapter 3. This is accom-

plished through the aggregation of quantitative transcription factor binding data matched to allelic sequence. We can leverage these pre-existing transcription factor binding sites containing one or more in silico "variants" to generate predictions of the consequences variations at every possible site along a binding motif has to transcription factor binding. SEMpl surpassed the current standard, position weight matrices (PWMs), when predicting measures of transcription factor binding from ChIP-seq data. SEMpl was able to recapitulate experimental measures of transcription factor binding, and outperformed other variant prediction methods. We found SEMpl to be cell-type agnostic, suggesting that nucleotide sequence alone does not drive differential transcription factor binding between cell-types.

I expanded the scope of SEMpl in Chapter 4 by including predictions of DNA methylation consequences on transcription factor binding sites. This computational tool, named SEMpl with Methylation (SEMplMe), is able to provide quantitative predictions of the effect of methylation on transcription factor binding affinity by following the same schema as SEMpl, with the addition of quantitative measures of DNA methylation genome-wide. This allows us to add two additional letters to the nucleotide alphabet, M – methylated cytosine, and W – guanine on the opposite strand of a methylated cytosine. SEMplMe agrees with known annotations of methylation sensitive and insensitive transcription factors, recapitulates prior experimental measures of transcription factor binding, and outperforms another method that predicts the methylation consequences in transcription factor binding sites. Unlike SEMpl, SEMplMe does differ between cell-types for methylated nucleotides, suggesting that DNA methylation is able to drive differential transcription factor binding in a cell-type specific manner.

Both SEMpl and SEMplMe are limited to currently available datasets. In order

to run both tools for a single transcription factor, a PWM, ChIP-seq data set, cell-type matched DNase-seq, and whole genome bisulfite sequencing data are required. Prediction confidence at each base is correlated to the number of instances of the in silico "variant" observed within the genome, which can be a limiting step in some cases of methylated sequences with few binding sites. It is also expected that some of our predictions made from genome-wide level annotations will not fully recapitulate endogenous locus-specific binding patterns due to factors external to the base sequence and methylation, such as cofactor binding.

Together SEMpl and SEMplMe can be used to prioritize non-coding variation for experimental follow-up. For instance, GWAS variants from the National Human Genome Research Institute (NHGRI) catalogue can be screened for overlap with predicted transcription factor binding sites by RegulomeDB, followed by ranking potential functional variants using SEMpl and SEMplMe [24, 36]. Experimental follow-up could be carried out based on different criteria, including variants most likely to disrupt endogenous transcription factor binding, sites predicted to bind to a specific transcription factor, or by regions of the genome known to be associated with a specific disease. To make predictions as accessible as possible, pre-computed matrices are available for >90 transcription factors from SEMpl, and >70 transcription factors from SEMplMe. Providing SEMpl and SEMplMe scores alongside RegulomeDB predictions of function can vastly improve the accessibility of these tools. The further expansion of this methodology through the addition of more functional annotations, such as hydroxymethylation and nonCpG methylation, may also help improve SEMplMe predictions [216, 142].

7.2 Characterizing the *in vivo* Activity of Regulatory Elements

The research presented in Chapters 5-6 describes novel experimental tools to validate and characterize regulatory elements in zebrafish transgenic models. Zebrafish is a classic model organism for the study of genomic regulatory elements and gene function. In Chapter 5 I demonstrate that a widely used experimental method for modifying gene expression is also capable of functioning in a zebrafish cell line. The *lac* operator-repressor system is an inducible repression system where an expressed protein (LacI) specifically binds to its operator sequence (*lacO*). When *lacO* sites are placed between a promoter and a gene, the binding of LacI is able to prevent expression of the gene via steric hindrance [161]. This repression system is often used in mice and human cell lines to study gene activation and repression. I demonstrate functionality of this system in zebrafish cells by repressing the expression of a luciferase reporter gene in the PAC2 zebrafish fibroblast cell line. Luciferase expression was rescued with the addition of the allosteric inhibitor Isopropyl β-d-1-thiogalactopyranoside (IPTG), which prevents LacI binding, providing evidence that the observed reduction of luciferase signal was due to the *lac* operator-repressor system. The validation of this system in zebrafish cells strongly suggests it is functional in whole zebrafish, providing a new way to investigate gene expression using this model organism.

Following the validation of the *lac* operator-repressor system in zebrafish cells, I utilized this repression system and a dCas9 repression system as the basis of two novel reporter assays to characterize the function of negative regulatory elements in transgenic zebrafish. Negative regulatory elements, such as silencers and enhancer blockers, are associated with a decrease in gene expression and are largely understud-

ied due to challenges detecting their regulatory activity. In Chapter 6, I introduce two inversion assays that are able to take a negative regulatory element signal and invert it into positive reporter expression. The dCas9 inversion assay uses negative regulatory activity to decrease the expression of an sgRNA, reducing dCas9 binding and subsequent blocking of GFP reporter gene transcription. The *lac* operator-repressor inversion assay uses negative regulatory activity to decrease the expression of LacI, reducing its binding to *lacO* sites located downstream of a promoter driving GFP reporter expression. The positive reporter output of these assays represents a breakthrough in the study of negative regulatory elements. These assays will allow the spatio-temporal characterization of negative regulatory elements *in vivo* in whole animals, a feat previously only possible for positive regulatory elements, such as promoters and enhancers. While these assays are promising, it remains to be seen if they can work as reporter assays for negative regulatory elements in transgenic animal models. Early attempts to utilize the dCas9 inversion assay in whole zebrafish resulted in highly mosaic animals, potentially due to the size of the plasmid that integrates into the genome. While the feasibility of the *lac* operator-repressor system is promising, this inversion assay has yet to be tested in zebrafish. Reporter assays in general are limited in their construction, as they may not recapitulate endogenous patterns of DNA methylation or histone modifications. There is possible bias against tested regulatory elements, such as those required to be a specific distance from their target gene, or are orientation or promoter-dependent. This assay is unlikely to be able to identify elements with weak regulatory activity. Future applications of these inversion assays allow for the unbiased identification of novel negative regulatory elements. This includes a potential silencer trap, where our assay is randomly integrated into the zebrafish genome without the addition of a putative

silencer element. Nearby silencers will then be able to activate reporter gene expression. The dCas9 inversion assay can be used to identify negative regulatory elements in a high-throughput manner, by employing the sgRNA recognition sequence as a barcode. This would allow negative regulatory elements to specifically repress only the reporter activity expressed from their plasmid of origin. Following fluorescent cell sorting and sequencing of GFP positive cells, we would be able to use this assay to discover novel negative regulatory elements in a high-throughput fashion.

7.3 Concluding Remarks

The work in this thesis has provided novel computational and experimental tools focused on elucidating the function of non-coding regulatory sequence in the human genome. These tools are expected to provide a new baseline from which regulatory elements can be validated and their function characterized. They will allow for spatio-temporal characterization of entire regulatory elements, and investigation of the transcription factor binding sites within them down to the nucleotide level.

While this work is critical to further the understanding of non-coding genomic sequences, additional experimental validations of these sequences will be required to fully elucidate the non-coding human genome. Predictions of functional non-coding variants provided by SEMpl and SEMplMe will require experimental validation by reporter assay, CRISPR-based mutagenesis, or *in vitro* methods such as EMSA, to establish their putative roles in altering transcription factor binding affinity. Methods such as RNA-seq of mutagenized cell culture can help establish these variant's effects on global gene expression. Validating negative regulatory elements by inverted reporter assays is especially important now that high-throughput technologies to

identify putative silencer activity are being released [180, 187]. In addition to characterizing tissue-specific regulatory activity, methods such as chromatin conformation capture technologies can be implemented for validated negative regulatory elements to identify their target genes. This work is expected to further unravel genome-wide regulatory networks and elucidate non-coding loci capable of contributing to human disease.

Going forward the prediction of non-coding variation and discovery of non-coding regulatory elements will help to form a foundation toward the understanding of the non-coding regulatory genome. Generating tools such as the VISTA Enhancer Browser for silencers and genome-wide maps of insulators will allow researchers to make more in depth predictions of function and better capture the complexities of gene regulatory landscapes [31]. Furthermore, increasing understanding of non-coding regions of the genome will bolster the use of whole-genome sequencing methods. The widespread use of these methods will help us to better understand sequencing variance in non-coding regions, assisting in predictions of disease-causing variation in non-coding regions genome-wide. Eventually a firm understanding of non-coding variation and regulatory networks will be able to contribute to healthcare outcomes by expanding our understanding of gene regulation in personalized medicine and key non-coding variants contributing to rare disease.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[2] Maya Kasowski, Fabian Grubert, Christopher Heffelfinger, Manoj Hariharan, Akwasi Asabere, Sebastian M. Waszak, Lukas Habegger, Joel Rozowsky, Minyi Shi, Alexander E. Urban, and et al. Variation in transcription factor binding among humans. *Science*, 328(5975):232–235, Apr 2010.

[3] Lucia A. Hindorff, Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–9367, Jun 2009.

[4] Alexander Gusev, S. Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J. Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, and et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American journal of human genetics*, 95(5):535–552, Nov 2014.

[5] Robin van der Lee, Robin van der Lee, Solenne Correard, and Wyeth W. Wasserman. Deregulated regulators: Disease-causing cis variants in transcription factor genes. *Trends in Genetics*, 36(7):523–539, 2020.

[6] Frederick Kinyua Kamanu, Yulia A. Medvedeva, Ulf Schaefer, Boris R. Jankovic, John A. C. Archer, and Vladimir B. Bajic. Mutations and binding sites of human transcription factors. *Frontiers in genetics*, 3:100, Jun 2012.

[7] Eileen Sproat Emison, Andrew S. McCallion, Carl S. Kashuk, Richard T. Bush, Elizabeth Grice, Shin Lin, Matthew E. Portnoy, David J. Cutler, Eric D. Green, and Aravinda Chakravarti. A common sex-dependent mutation in a ret enhancer underlies hirschsprung disease risk. *Nature*, 434(7035):857–863, Apr 2005.

[8] Kyle J. Gaulton, Takao Nammo, Lorenzo Pasquali, Jeremy M. Simon, Paul G. Giresi, Marie P. Fogarty, Tami M. Panhuis, Piotr Mieczkowski, Antonio Secchi, Domenico Bosco, and et al. A map of open chromatin in human pancreatic islets. *Nature genetics*, 42(3):255–259, Mar 2010.

[9] Fedik Rahimov, Mary L. Marazita, Axel Visel, Margaret E. Cooper, Michael J. Hitchler, Michele Rubini, Frederick E. Domann, Manika Govil, Kaare Christensen, Camille Bille, and et al. Disruption of an ap-2alpha binding site in an irf6 enhancer is associated with cleft lip. *Nature genetics*, 40(11):1341–1347, Nov 2008.

[10] J. M. Heckmann, H. Uwimpuhwe, R. Ballo, M. Kaur, V. B. Bajic, and S. Prince. A functional snp in the regulatory region of the decay-accelerating factor gene associates with extraocular muscle pareses in myasthenia gravis. *Genes and immunity*, 11(1):1–10, Jan 2010.

[11] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E. Lee, Tim Ahfeldt, Katherine V. Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M. Ruda, and et al. From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, Aug 2010.

[12] Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, and et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, May 2015.

[13] Erich Roessler, Ping Hu, Sung-Kook Hong, Kshitij Srivastava, Blake Carrington, Raman Sood, Hanna Petrykowska, Laura Elnitski, Lucilene A. Ribeiro, Antonio Richieri-Costa, and et al. Unique alterations of an ultraconserved non-coding element in the 3'utr of zic2 in holoprosencephaly. *PloS one*, 7(7):e39026, Jul 2012.

[14] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, Feb 2012.

[15] Adam G. Diehl and Alan P. Boyle. Deciphering encode. *Trends in genetics: TIG*, 32(4):238–249, Apr 2016.

[16] Dimitris Polychronopoulos, James W. D. King, Alexander J. Nash, Ge Tan, and Boris Lenhard. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic acids research*, 45(22):12611–12624, Dec 2017.

[17] Nadav Ahituv, Yiwen Zhu, Axel Visel, Amy Holt, Veena Afzal, Len A. Pennacchio, and Edward M. Rubin. Deletion of ultraconserved elements yields viable mice. *PLoS biology*, 5(9):e234, Sep 2007.

[18] The 1000 Genomes Project Consortium and The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[19] Melissa J. Fullwood and Yijun Ruan. Chip-based methods for the identification of long-range chromatin interactions. *Journal of cellular biochemistry*, 107(1):30–39, May 2009.

[20] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, and et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

[21] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[22] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620):20120362, May 2013.

[23] Sierra S. Nishizaki and Alan P. Boyle. Mining the unknown: Assigning function to noncoding single nucleotide polymorphisms. *Trends in genetics: TIG*, 33(1):34–45, Jan 2017.

[24] Alan P. Boyle, Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, Julie Park, Benjamin C. Hitz, Shuai Weng, and et al. Annotation of functional variation in personal genomes using regulomedb. *Genome research*, 22(9):1790–1797, Sep 2012.

[25] Shengcheng Dong and Alan P. Boyle. Predicting functional variants in enhancer and promoter elements using regulomedb. *Human Mutation*, 40(9):1292–1298, 2019.

[26] Tianyin Zhou, Ning Shen, Lin Yang, Namiko Abe, John Horton, Richard S. Mann, Harmen J. Bussemaker, Raluca Gordân, and Remo Rohs. Quantitative modeling of transcription factor binding specificities using dna shape. *Proceedings of the National Academy of Sciences of the United States of America*, 112(15):4654–4659, Apr 2015.

[27] Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, Mar 2014.

[28] Dongwon Lee, David U. Gorkin, Maggie Baker, Benjamin J. Strober, Alessandro L. Asoni, Andrew S. McCallion, and Michael A. Beer. A method to predict the impact of regulatory variants from dna sequence. *Nature genetics*, 47(8):955–961, Aug 2015.

[29] Shannon Fisher, Elizabeth A. Grice, Ryan M. Vinton, Seneca L. Bessling, Akihiro Urasaki, Koichi Kawakami, and Andrew S. McCallion. Evaluating the biological relevance of putative enhancers using tol2 transposon-mediated transgenesis in zebrafish. *Nature protocols*, 1(3):1297–1305, 2006.

[30] Len A. Pennacchio, Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, Keith D. Lewis, and et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, Nov 2006.

[31] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A. Pennacchio. Vista enhancer browser–a database of tissue-specific human enhancers. *Nucleic acids research*, 35(Database issue):D88–92, Jan 2007.

[32] Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research*, 23(5):800–811, May 2013.

[33] Cosmas D. Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by starr-seq. *Science*, 339(6123):1074–1077, Mar 2013.

[34] Hanna M. Petrykowska, Christopher M. Vockley, and Laura Elnitski. Detection and characterization of silencers and enhancer-blockers in the greater cftr locus. *Genome research*, 18(8):1238–1246, Aug 2008.

[35] Mark M. Pomerantz, Nasim Ahmadiyeh, Li Jia, Paula Herman, Michael P. Verzi, Harshavardhan Doddapaneni, Christine A. Beckwith, Jennifer A. Chan, Adam Hills, Matt Davis, and et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with myc in colorectal cancer. *Nature genetics*, 41(8):882–884, Aug 2009.

[36] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(Database issue):D1001–6, Jan 2014.

[37] Wellcome Trust Case Control Consortium, Julian B. Maller, Gilean McVean, Jake Byrnes, Damjan Vukcevic, Kimmo Palin, Zhan Su, Joanna M. M. Howson, Adam Auton, Simon Myers, and et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, 44(12):1294–1301, Dec 2012.

[38] Daniel E. Bauer, Sophia C. Kamran, Samuel Lessard, Jian Xu, Yuko Fujiwara, Carrie Lin, Zhen Shao, Matthew C. Canver, Elenoe C. Smith, Luca Pinello, and et al. An erythroid enhancer of bcl11a subject to genetic variation determines fetal hemoglobin level. *Science*, 342(6155):253–257, Oct 2013.

[39] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke, and Zlatko Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2):256–278, Mar 2014.

[40] Graham R. S. Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature methods*, 11(3):294–296, Mar 2014.

[41] Ryan McDaniell, Bum-Kyu Lee, Lingyun Song, Zheng Liu, Alan P. Boyle, Michael R. Erdos, Laura J. Scott, Mario A. Morken, Katerina S. Kucera, Anna Battenhouse, and et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, 328(5975):235–239, Apr 2010.

[42] Jacob F. Degner, Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E. Crawford, and et al. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482(7385):390–394, Feb 2012.

[43] T. S. Furey and P. Sethupathy. Genetics driving epigenetics. *Science*, 342(6159):705–706, 2013.

[44] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, and et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, Sep 2012.

[45] Marc A. Schaub, Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9):1748–1759, Sep 2012.

[46] Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W. Wasserman. Dna shape features improve transcription factor binding site predictions in vivo. *Cell systems*, 3(3):278–286.e4, Sep 2016.

[47] Simon G. Coetzee, Suhn K. Rhie, Benjamin P. Berman, Gerhard A. Coetzee, and Houtan Noushmehr. Funcisnp: an r/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory snps. *Nucleic Acids Research*, 40(18):e139–e139, 2012.

[48] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics*, 26(16):2069–2070, 2010.

[49] Lucas D. Ward and Manolis Kellis. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 337(6102):1675–1678, Sep 2012.

[50] Brad Gulko, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature genetics*, 47(3):276–283, Mar 2015.

[51] Mulin Jun Li, Lily Yan Wang, Zhengyuan Xia, Pak Chung Sham, and Junwen Wang. Gwas3d: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Research*, 41(W1):W150–W158, 2013.

[52] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164, Sep 2010.

[53] Lucas D. Ward and Manolis Kellis. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, 40(Database issue):D930–4, Jan 2012.

[54] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[55] Sebastian Okser, Tapio Pahikkala, Antti Airola, Tapio Salakoski, Samuli Ripatti, and Tero Aittokallio. Regularized machine learning in the genetic prediction of complex traits. *PLoS genetics*, 10(11):e1004754, Nov 2014.

[56] Daniel Quang, Yifei Chen, and Xiaohui Xie. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, 2015.

[57] Hashem A. Shihab, Mark F. Rogers, Julian Gough, Matthew Mort, David N. Cooper, Ian N. M. Day, Tom R. Gaunt, and Colin Campbell. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10):1536–1543, May 2015.

[58] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10):931–934, Oct 2015.

[59] Peter D. Stenson, Matthew Mort, Edward V. Ball, Katy Shaw, Andrew Phillips, and David N. Cooper. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genetics*, 133(1):1–9, Jan 2014.

[60] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 2009.

[61] Gerald A. Higgins, Ari Allyn-Feuer, and Brian D. Athey. Epigenomic mapping and effect sizes of noncoding variants associated with psychotropic drug response. *Pharmacogenomics*, 16(14):1565–1583, Sep 2015.

[62] Huiling He, Wei Li, Sandya Liyanarachchi, Mukund Srinivas, Yanqiang Wang, Keiko Akagi, Yao Wang, Dayong Wu, Qianben Wang, Victor Jin, and et al. Multiple functional variants in long-range enhancer elements contribute to the risk of snp rs965513 in thyroid cancer. *Proceedings of the National Academy of Sciences*, 112(19):6128–6133, 2015.

[63] Rupali P. Patwardhan, Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, Jennifer M. Andrie, Su-In Lee, Gregory M. Cooper, and et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*, 30(3):265–270, Feb 2012.

[64] Davis J. McCarthy, Peter Humburg, Alexander Kanapin, Manuel A. Rivas, Kyle Gaulton, Jean-Baptiste Cazier, Peter Donnelly, and asds. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3):26, 2014.

[65] Norihiro Kato, Marie Loh, Fumihiko Takeuchi, Niek Verweij, Xu Wang, Weihua Zhang, Tanika N. Kelly, Danish Saleheen, Benjamin Lehne, Irene Mateo Leach, and et al. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for dna methylation. *Nature genetics*, 47(11):1282–1293, Nov 2015.

[66] Matthew H. Law, D. Timothy Bishop, Jeffrey E. Lee, Myriam Brossard, Nicholas G. Martin, Eric K. Moses, Fengju Song, Jennifer H. Barrett, Rajiv Kumar, Douglas F. Easton, and et al. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nature genetics*, 47(9):987–995, Sep 2015.

[67] Carlo Sidore, Fabio Busonero, Andrea Maschio, Eleonora Porcu, Silvia Naitza, Magdalena Zoledziewska, Antonella Mulas, Giorgio Pistis, Maristella Steri, Fabrice Danjou, and et al. Genome sequencing elucidates sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature genetics*, 47(11):1272–1281, Nov 2015.

[68] Geng Chen, Dianke Yu, Jiwei Chen, Ruifang Cao, Juan Yang, Huan Wang, Xiangjun Ji, Baitang Ning, and Tieliu Shi. Re-annotation of presumed noncoding disease/trait-associated genetic variants by integrative analyses. *Scientific reports*, 5:9453, Mar 2015.

[69] Stacey L. Edwards, Jonathan Beesley, Juliet D. French, and Alison M. Dunning. Beyond gwass: Illuminating the dark road from association to function. *The American Journal of Human Genetics*, 93(5):779–797, 2013.

[70] Matthew L. Freedman, Alvaro N. A. Monteiro, Simon A. Gayther, Gerhard A. Coetzee, Angela Risch, Christoph Plass, Graham Casey, Mariella De Biasi, Chris Carlson, David Duggan, and et al. Principles for the post-gwas functional characterization of cancer risk loci. *Nature Genetics*, 43(6):513–518, 2011.

[71] M. G. Fried. Measurement of protein-dna interaction parameters by electrophoresis mobility shift assay. *Electrophoresis*, 10(5-6):366–376, May 1989.

[72] Falk Butter, Lucy Davison, Tar Viturawong, Marion Scheibe, Michiel Vermeulen, John A. Todd, and Matthias Mann. Proteome-wide analysis of disease-associated snps that show allele-specific transcription factor binding. *PLoS Genetics*, 8(9):e1002982, 2012.

[73] Nathaniel D. Heintzman and Bing Ren. Finding distal regulatory elements in the human genome. *Current Opinion in Genetics Development*, 19(6):541–549, 2009.

[74] James O. J. Davies, Jelena M. Telenius, Simon J. McGowan, Nigel A. Roberts, Stephen Taylor, Douglas R. Higgs, and Jim R. Hughes. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature Methods*, 13(1):74–80, 2016.

[75] Hélène Hagège, Petra Klous, Caroline Braem, Erik Splinter, Job Dekker, Guy Cathala, Wouter de Laat, and Thierry Forné. Quantitative analysis of chromosome conformation capture assays (3c-qpcr). *Nature protocols*, 2(7):1722–1733, 2007.

[76] Borbala Mifsud, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, Simon Andrews, William Grey, Philip A. Ewels, and et al. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nature genetics*, 47(6):598–606, Jun 2015.

[77] E. de Wit, E. de Wit, and W. de Laat. A decade of 3c technologies: insights into nuclear organization. *Genes Development*, 26(1):11–24, 2012.

[78] Joerg Ermann and Laurie H. Glimcher. After gwas: mice to the rescue? *Current opinion in immunology*, 24(5):564–570, Oct 2012.

[79] Feng Zhang and James R. Lupski. Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1):R102–10, Oct 2015.

[80] Olivia Corradin, Alina Saiakhova, Batool Akhtar-Zaidi, Lois Myeroff, Joseph Willis, Richard Cowper-Sal lari, Mathieu Lupien, Sanford Markowitz, and Peter C. Scacheri. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome research*, 24(1):1–13, Jan 2014.

[81] Jenny C. Taylor, Hilary C. Martin, Stefano Lise, John Broxholme, Jean-Baptiste Cazier, Andy Rimmer, Alexander Kanapin, Gerton Lunter, Simon Fiddy, Chris Allan, and et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature genetics*, 47(7):717–726, Jul 2015.

[82] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, Jan 2001.

[83] Julia E. VanderMeer and Nadav Ahituv. cis-regulatory mutations are a genetic cause of human limb malformations. *Developmental Dynamics*, 240(5):920–930, 2011.

[84] Marie P. Fogarty, Maren E. Cannon, Swarooparani Vadlamudi, Kyle J. Gaulton, and Karen L. Mohlke. Identification of a regulatory variant that binds foxa1 and foxa2 at the cdc123/camk1d type 2 diabetes gwas locus. *PLoS genetics*, 10(9):e1004633, Sep 2014.

[85] Daniel Savic, Honggang Ye, Ivy Aneas, Soo-Young Park, Graeme I. Bell, and Marcelo A. Nobrega. Alterations in tcf7l2 expression define its role as a key regulator of glucose metabolism. *Genome research*, 21(9):1417–1425, Sep 2011.

[86] Michael L. Stitzel, Praveen Sethupathy, Daniel S. Pearson, Peter S. Chines, Lingyun Song, Michael R. Erdos, Ryan Welch, Stephen C. J. Parker, Alan P. Boyle, Laura J. Scott, and et al. Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metabolism*, 12(5):443–455, 2010.

[87] Husen M. Umer, Marco Cavalli, Michal J. Dabrowski, Klev Diamanti, Marcin Kruczyk, Gang Pan, Jan Komorowski, and Claes Wadelius. A significant regulatory mutation burden at a high-affinity position of the ctcf motif in gastrointestinal cancers. *Human Mutation*, 37(9):904–913, 2016.

[88] Gary D. Stormo, Thomas D. Schneider, Larry Gold, and Andrzej Ehrenfeucht. Use of the "perceptron" algorithm to distinguish translational initiation sites ine. coli. *Nucleic Acids Research*, 10(9):2997–3011, 1982.

[89] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, and et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 31(2):126–134, Feb 2013.

[90] Malin C. Andersen, Pär G. Engström, Stuart Lithwick, David Arenillas, Per Eriksson, Boris Lenhard, Wyeth W. Wasserman, and Jacob Odeberg. In silico detection of sequence variations modifying transcriptional regulation. *PLoS Computational Biology*, 4(1):e5, 2008.

[91] Maxim Barenboim and Thomas Manke. Chromos: an integrated web tool for snp classification, prioritization and functional interpretation. *Bioinformatics*, 29(17):2197–2198, Sep 2013.

[92] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A. Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R. Kulkarni, Ge Tan, and et al. Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, 46(D1):D1284, Jan 2018.

[93] Geoff Macintyre, James Bailey, Izhak Haviv, and Adam Kowalczyk. is-rsnp: a novel technique for in silico regulatory snp detection. *Bioinformatics*, 26(18):i524–30, Sep 2010.

[94] Thomas Manke, Matthias Heinig, and Martin Vingron. Quantifying the effect of sequence variation on regulatory interactions. *Human Mutation*, 31(4):477–483, 2010.

[95] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *bioRxiv*, 2017.

[96] Ilya E. Vorontsov, Ivan V. Kulakovskiy, Grigory Khimulya, Daria D. Nikolaeva, and Vsevolod J. Makeev. Perfectos-ape - predicting regulatory functional effect of snps by approximate p-value estimation. *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, 2015.

[97] Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, Aug 2015.

[98] B. C. Foat, A. V. Morozov, and H. J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, 22(14):e141–e149, 2006.

[99] A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpaa, and et al. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome Research*, 20(6):861–873, 2010.

[100] Todd R. Riley, Allan Lazarovici, Richard S. Mann, and Harmen J. Bussemaker. Building accurate sequence-to-affinity models from high-throughput in vitro protein-dna binding data using featurereduce. *eLife*, 4, Dec 2015.

[101] Yue Zhao and Gary D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology*, 29(6):480–483, Jun 2011.

[102] Richard Cowper-Sal lari, Xiaoyang Zhang, Jason B. Wright, Swneke D. Bailey, Michael D. Cole, Jerome Eeckhoute, Jason H. Moore, and Mathieu Lupien. Breast cancer risk-associated snps modulate the affinity of chromatin for foxa1 and alter gene expression. *Nature genetics*, 44(11):1191–1198, Nov 2012.

[103] Maxwell A. Hume, Luis A. Barrera, Stephen S. Gisselbrecht, and Martha L. Bulyk. Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein-dna interactions. *Nucleic acids research*, 43(Database issue):D117–22, Jan 2015.

[104] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, and et al. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.

[105] Hélène Touzet and Jean-Stéphane Varré. Efficient and accurate p-value computation for position weight matrices. *Algorithms for molecular biology: AMB*, 2:15, Dec 2007.

[106] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, Mar 2009.

[107] Oliver Bembom. Sequence logos for dna sequence alignments. 2018.

[108] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

[109] Jieming Chen, Joel Rozowsky, Timur R. Galeev, Arif Harmanci, Robert Kitchen, Jason Bedford, Alexej Abyzov, Yong Kong, Lynne Regan, and Mark Gerstein. A uniform survey of allele-specific binding and expression over 1000-genomes-project individuals. *Nature communications*, 7:11101, Apr 2016.

[110] Zheng Zuo, Basab Roy, Yiming Kenny Chang, David Granas, and Gary D. Stormo. Measuring quantitative effects of methylation on transcription factor–dna binding affinity. *Science Advances*, 3(11):eaao1799, 2017.

[111] Victor G. Levitsky, Ivan V. Kulakovskiy, Nikita I. Ershov, Dmitry Oshchepkov, Vsevolod J. Makeev, T. C. Hodgman, and Tatyana I. Merkulova. Application of experimentally verified transcription factor binding sites models for computational analysis of chip-seq data. *BMC Genomics*, 15(1):80, 2014.

[112] Nilesh Aghera, Ninganna Earanna, and Jayant B. Udgaonkar. Equilibrium unfolding studies of monellin: The double-chain variant appears to be more stable than the single-chain variant. *Biochemistry*, 50(13):2434–2444, 2011.

[113] Caroline A. Schneider, Wayne S. Rasband, and Kevin W. Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671–675, Jul 2012.

[114] David Cappellen, Thomas Schlange, Matthieu Bauer, Francisca Maurer, and Nancy E. Hynes. Novel c-myc target genes mediate differential effects on cell proliferation and migration. *EMBO reports*, 8(1):70–76, Jan 2007.

[115] Swneke D. Bailey, Xiaoyang Zhang, Kinjal Desai, Malika Aid, Olivia Corradin, Richard Cowper-Sal·lari, Batool Akhtar-Zaidi, Peter C. Scacheri, Benjamin Haibe-Kains, and Mathieu Lupien. Znf143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature Communications*, 6(1), 2015.

[116] Jiangchuan Ye, Nathan R. Tucker, Lu-Chen Weng, Sebastian Clauss, Steven A. Lubitz, and Patrick T. Ellinor. A functional variant associated with atrial fibrillation regulates pitx2c expression through tfap2a. *The American Journal of Human Genetics*, 99(6):1281–1291, 2016.

[117] Huoru Zhang, Yan Zhang, Yong-Fei Wang, David Morris, Nattiya Hirankarn, Yujun Sheng, Jiangshan Shen, Hai-Feng Pan, Jing Yang, Sen Yang, and et al. Meta-analysis of gwas on both chinese and european populations identifies gpr173 as a novel x chromosome susceptibility gene for sle. *Arthritis Research Therapy*, 20(1), 2018.

[118] Sierra S. Nishizaki, Natalie Ng, Shengcheng Dong, Robert S. Porter, Cody Morterud, Colten Williams, Courtney Asman, Jessica A. Switzenberg, and Alan P. Boyle. Predicting the effects of snps on transcription factor binding affinity. *Bioinformatics*, 36(2):364–372, Jan 2020.

[119] A. P. Bird. Cpg-rich islands and the function of dna methylation. *Nature*, 321(6067):209–213, 1986.

[120] P. H. Tate and A. P. Bird. Effects of dna methylation on dna-binding proteins and gene expression. *Current opinion in genetics development*, 3(2):226–231, Apr 1993.

[121] Matthew T. Maurano, Hao Wang, Sam John, Anthony Shafer, Theresa Canfield, Kristen Lee, and John A. Stamatoyannopoulos. Role of dna methylation in modulating transcription factor occupancy. *Cell reports*, 12(7):1184–1195, Aug 2015.

[122] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K. Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, and et al. Impact of cytosine methylation on dna binding specificities of human transcription factors. *Science*, 356(6337), May 2017.

[123] David P. Gavin and Rajiv P. Sharma. Histone modifications, dna methylation, and schizophrenia. *Neuroscience and biobehavioral reviews*, 34(6):882–888, May 2010.

[124] Yong-Hui Jiang, Trilochan Sahoo, Ron C. Michaelis, Dani Bercovich, Jan Bressler, Catherine D. Kashork, Qian Liu, Lisa G. Shaffer, Richard J. Schroer, David W. Stockton, and et al. A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for ube3a. *American journal of medical genetics. Part A*, 131(1):1–10, Nov 2004.

[125] Keith D. Robertson. Dna methylation and human disease. *Nature reviews. Genetics*, 6(8):597–610, Aug 2005.

[126] Shaohui Hu, Jun Wan, Yijing Su, Qifeng Song, Yaxue Zeng, Ha Nam Nguyen, Jaehoon Shin, Eric Cox, Hee Sool Rho, Crystal Woodard, and et al. Dna methylation presents distinct binding sites for human transcription factors. *eLife*, 2, 2013.

[127] Ishminder K. Mann, Raghunath Chatterjee, Jianfei Zhao, Ximiao He, Matthew T. Weirauch, Timothy R. Hughes, and Charles Vinson. Cg methylated microarrays identify a novel methylated sequence bound by the cebpb—atf4 heterodimer that is active in vivo. *Genome research*, 23(6):988–997, Jun 2013.

[128] Desiree Tillo, Sreejana Ray, Khund-Sayeed Syed, Mary Rose Gaylor, Ximiao He, Jun Wang, Nima Assad, Stewart R. Durell, Aleksey Porollo, Matthew T. Weirauch, and et al. The epstein-barr virus b-zip protein zta recognizes specific dna sequences containing 5-methylcytosine and 5-hydroxymethylcytosine. *Biochemistry*, 56(47):6200–6210, Nov 2017.

[129] Heng Zhu, Guohua Wang, and Jiang Qian. Transcription factors as readers and effectors of dna methylation. *Nature reviews. Genetics*, 17(9):551–565, Aug 2016.

[130] Judith F. Kribelbauer, Oleg Laptenko, Siying Chen, Gabriella D. Martini, William A. Freed-Pastor, Carol Prives, Richard S. Mann, and Harmen J. Bussemaker. Quantitative analysis of the dna methylation sensitivity of transcription factor complexes. *Cell reports*, 19(11):2383–2395, Jun 2017.

[131] Guohua Wang, Ximei Luo, Jianan Wang, Jun Wan, Shuli Xia, Heng Zhu, Jiang Qian, and Yadong Wang. Medreaders: a database for transcription factors that bind to methylated dna. *Nucleic acids research*, 46(D1):D146–D151, Jan 2018.

[132] Quy Xiao Xuan Lin, Stephanie Sian, Omer An, Denis Thieffry, Sudhakar Jha, and Touati Benoukraf. Methmotif: an integrative cell specific database of transcription factor binding motifs coupled with dna methylation profiles. *Nucleic acids research*, 47(D1):D145–D154, Jan 2019.

[133] Michael B. Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, Christiane Wirbelauer, Edward J. Oakeley, Dimos Gaidatzis, and et al. Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, Dec 2011.

[134] Yi-Lan Weng, Ran An, Jaehoon Shin, Hongjun Song, and Guo-Li Ming. Dna modifications and neurological disorders. *Neurotherapeutics: the journal of the American Society for Experimental NeuroTherapeutics*, 10(4):556–567, Oct 2013.

[135] Till Bartke, Michiel Vermeulen, Blerta Xhemalce, Samuel C. Robson, Matthias Mann, and Tony Kouzarides. Nucleosome-interacting proteins regulated by dna and histone methylation. *Cell*, 143(3):470–484, Oct 2010.

[136] Donghong Zhang, Bingruo Wu, Ping Wang, Yidong Wang, Pengfei Lu, Tamilla Nechiporuk, Thomas Floss, John M. Greally, Deyou Zheng, and Bin Zhou. Non-cpg methylation by dnmt3b facilitates rest binding and gene silencing in developing mouse hearts. *Nucleic acids research*, 45(6):3102–3115, Apr 2017.

[137] Jonathan R. Moll, Asha Acharya, Jozsef Gal, Alain A. Mir, and Charles Vinson. Magnesium is required for specific dna binding of the creb b-zip domain. *Nucleic acids research*, 30(5):1240–1246, Mar 2002.

[138] A. C. Bell and G. Felsenfeld. Methylation of a ctcf-dependent boundary controls imprinted expression of the igf2 gene. *Nature*, 405(6785):482–485, May 2000.

[139] M. P. Stadnick, F. M. Pieracci, M. J. Cranston, E. Taksel, J. L. Thorvaldsen, and M. S. Bartolomei. Role of a 461-bp g-rich repetitive element in h19 transgene imprinting. *Development genes and evolution*, 209(4):239–248, Apr 1999.

[140] Mario Renda, Ilaria Baglivo, Bonnie Burgess-Beusse, Sabrina Esposito, Roberto Fattorusso, Gary Felsenfeld, and Paolo V. Pedone. Critical dna binding interactions of the insulator protein ctcf: a small number of zinc fingers mediate strong binding, and a single finger-dna interaction controls binding at imprinted loci. *The Journal of biological chemistry*, 282(46):33336–33345, Nov 2007.

[141] L. Han, I. G. Lin, and C. L. Hsieh. Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. *Molecular and cellular biology*, 21(10):3416–3424, May 2001.

[142] Cornelia G. Spruijt, Felix Gnerlich, Arne H. Smits, Toni Pfaffeneder, Pascal W. T. C. Jansen, Christina Bauer, Martin Münzel, Mirko Wagner, Markus Müller, Fariha Khan, and et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, 152(5):1146–1159, Feb 2013.

[143] Xiaoji Wu and Yi Zhang. Tet-mediated active dna demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 18(9):517–534, 2017.

[144] Sierra S Nishizaki and Alan P Boyle. Semplme: A tool for integrating dna methylation effects in transcription factor binding affinity predictions. *bioRxiv*, 2020.

[145] David M. Langenau, Hui Feng, Stephane Berghmans, John P. Kanki, Jeffery L. Kutok, and A. Thomas Look. Cre/lox-regulated transgenic zebrafish model with conditional myc-induced t cell acute lymphoblastic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 102(17):6068–6073, Apr 2005.

[146] Stefan Hans, Jan Kaslin, Dorian Freudenreich, and Michael Brand. Temporally-controlled site-specific recombination in zebrafish. *PloS one*, 4(2):e4640, Feb 2009.

[147] Leah J. Campbell, John J. Willoughby, and Abbie M. Jensen. Two types of tet-on transgenic lines for doxycycline-inducible gene expression in zebrafish rod photoreceptors and a gateway-based tet-on toolkit. *PloS one*, 7(12):e51270, Dec 2012.

[148] Andrew Dodd, Stephen P. Chambers, and Donald R. Love. Short interfering rna-mediated gene targeting in the zebrafish. *FEBS letters*, 561(1-3):89–93, Mar 2004.

[149] Amanda Kelly and Adam F. Hurlstone. The use of rnai technologies for gene knockdown in zebrafish. *Briefings in functional genomics*, 10(4):189–196, Jul 2011.

[150] Woong Y. Hwang, Yanfang Fu, Deepak Reyon, Morgan L. Maeder, Prakriti Kaini, Jeffry D. Sander, J. Keith Joung, Randall T. Peterson, and Jing-Ruey Joanna Yeh. Heritable and precise zebrafish genome editing using a crispr-cas system. *PloS one*, 8(7):e68708, Jul 2013.

[151] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3:318–356, Jun 1961.

[152] Myles Brown, James Figge, Ulla Hansen, Christopher Wright, Kuan-Teh Jeang, George Khoury, David M. Livingston, and Thomas M. Roberts. Lac repressor can regulate expression from a hybrid sv40 early promoter containing a lac operator in animal cells. *Cell*, 49(5):603–612, 1987.

[153] A. Fieck, D. L. Wyborski, and J. M. Short. Modifications of the e.coli lac repressor for expression in eukaryotic cells: effects of nuclear signal sequences on protein activity and nuclear accumulation. *Nucleic acids research*, 20(7):1785–1791, Apr 1992.

[154] M. C. Hu and N. Davidson. The inducible lac operator-repressor system is functional in mammalian cells. *Cell*, 48(4):555–566, Feb 1987.

[155] H. S. Liu, E. S. Feliciano, and P. J. Stambrook. Cytochemical observation of regulated bacterial beta-galactosidase gene expression in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, 86(24):9951–9955, Dec 1989.

[156] J. Figge, C. Wright, C. J. Collins, T. M. Roberts, and D. M. Livingston. Stringent regulation of stably integrated chloramphenicol acetyl transferase genes by e. coli lac repressor in monkey cells. *Cell*, 52(5):713–722, Mar 1988.

[157] D. L. Wyborski, L. C. DuCoeur, and J. M. Short. Parameters affecting the use of the lac repressor system in eukaryotic cells and transgenic animals. *Environmental and molecular mutagenesis*, 28(4):447–458, 1996.

[158] Jessica L. Whited, Jessica A. Lehoczky, and Clifford J. Tabin. Inducible genetic system for the axolotl. *Proceedings of the National Academy of Sciences of the United States of America*, 109(34):13662–13667, Aug 2012.

[159] Seung-Hyun Myung, Junghee Park, Ji-Hye Han, and Tae-Hyoung Kim. Development of the mammalian expression vector system that can be induced by iptg and/or lactose. *Journal of microbiology and biotechnology*, 30(8):1124–1131, Aug 2020.

[160] Denis S. F. Biard, Michael R. James, André Cordier, and Alain Sarasin. Regulation of the escherichia coli lac operon expressed in human cells. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1130(1):68–74, 1992.

[161] Kwang-Ho Lee, Shirley Oghamian, Jin-A Park, Liang Kang, and Peter W. Laird. The remote-control system: a system for reversible and tunable control of endogenous gene expression in mice. *Nucleic acids research*, 45(21):12256–12269, Dec 2017.

[162] Nicole A. Vander Schaaf, Shirley Oghamian, Jin-A Park, Liang Kang, Peter W. Laird, and Kwang-Ho Lee. In vivo application of the remote-control system for the manipulation of endogenous gene expression. *Journal of visualized experiments: JoVE*, (145), Mar 2019.

[163] Vanessa Mella-Alvarado, Aude Gautier, Florence Le Gac, and Jean-Jacques Lareyre. Tissue and cell-specific transcriptional activity of the human cytomegalovirus immediate early gene promoter (ul123) in zebrafish. *Gene Expression Patterns*, 13(3-4):91–103, 2013.

[164] Zhenyu Jia, Jing Jia, Shuzhi Zhang, and Jiang Cao. Cmv enhancer may not be suitable for tissue-specific enhancement of promoters in cancer gene therapy. *Cancer gene therapy*, 27(5):389–392, May 2020.

[165] Tara L. Deans, Charles R. Cantor, and James J. Collins. A tunable genetic switch based on rnai and repressor proteins for regulating gene expression in mammalian cells. *Cell*, 130(2):363–372, Jul 2007.

[166] Luke A. Gilbert, Max A. Horlbeck, Britt Adamson, Jacqueline E. Villalta, Yuwen Chen, Evan H. Whitehead, Carla Guimaraes, Barbara Panning, Hidde L. Ploegh, Michael C. Bassik, and et al. Genome-scale crispr-mediated control of gene repression and activation. *Cell*, 159(3):647–661, Oct 2014.

[167] Koichi Kawakami. Tol2: a versatile gene transfer vector in vertebrates. *Genome biology*, 8 Suppl 1:S7, 2007.

[168] Andrea Martella, Mantas Matjusaitis, Jamie Auxillos, Steven M. Pollard, and Yizhi Cai. Emma: An extensible mammalian modular assembly toolkit for the rapid design and production of diverse expression vectors. *ACS synthetic biology*, 6(7):1380–1392, Jul 2017.

[169] Karen E. Gascoigne, Kozo Takeuchi, Aussie Suzuki, Tetsuya Hori, Tatsuo Fukagawa, and Iain M. Cheeseman. Induced ectopic kinetochore assembly bypasses the requirement for cenp-a nucleosomes. *Cell*, 145(3):410–422, Apr 2011.

[170] Niklas Senghaas and Reinhard W. Köster. Culturing and transfecting zebrafish pac2 fibroblast cells. *Cold Spring Harbor protocols*, 2009(6):db.prot5235, Jun 2009.

[171] Maximiliano L. Suster, Hiroshi Kikuta, Akihiro Urasaki, Kazuhide Asakawa, and Koichi Kawakami. Transgenesis in zebrafish with the tol2 transposon system. *Methods in molecular biology*, 561:41–63, 2009.

[172] Sierra S Nishizaki, Torrin L McDonald, Gregory A Farnum, Monica J Holmes, Melissa L Drexel, Jessica A Switzenberg, and Alan P Boyle. The inducible lac operator-repressor system is functional in zebrafish cells. *bioRxiv*, 2020.

[173] Jamie C. Kwasnieski, Christopher Fiore, Hemangi G. Chaudhari, and Barak A. Cohen. High-throughput functional testing of encode segmentation predictions. *Genome research*, 24(10):1595–1602, Oct 2014.

[174] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–825, Aug 2010.

[175] Michael M. Hoffman, Orion J. Buske, Jie Wang, Zhiping Weng, Jeff A. Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473–476, Mar 2012.

[176] S. Ogbourne and T. M. Antalis. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal*, 331 (Pt 1):1–14, Apr 1998.

[177] A. R. Clark and K. Docherty. Negative regulation of transcription in eukaryotes. *Biochemical Journal*, 296(3):521–541, 1993.

[178] J. M. Dong and L. Lim. The human neuronal alpha 1-chimaerin gene contains a position-dependent negative regulatory element in the first exon. *Neurochemical research*, 21(9):1023–1030, Sep 1996.

[179] Di Huang, Hanna M. Petrykowska, Brendan F. Miller, Laura Elnitski, and Ivan Ovcharenko. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome research*, 29(4):657–667, Apr 2019.

[180] Naresh Doni Jayavelu, Ajay Jajodia, Arpit Mishra, and R. David Hawkins. Candidate silencer elements for the human and mouse genomes. *Nature Communications*, 11(1), 2020.

[181] Stephen S. Gisselbrecht, Alexandre Palagi, Jesse V. Kurland, Julia M. Rogers, Hakan Ozadam, Ye Zhan, Job Dekker, and Martha L. Bulyk. Transcriptional silencers in drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. *Molecular cell*, 77(2):324–337.e8, Jan 2020.

[182] Xiangdong Lv, Zhijun Han, Hao Chen, Bo Yang, Xiaofeng Yang, Yuanxin Xia, Chenyu Pan, Lin Fu, Shuo Zhang, Hui Han, and et al. A positive role for polycomb in transcriptional regulation via h4k20me1. *Cell research*, 27(4):594, Apr 2017.

[183] Matthew D. Young, Tracy A. Willson, Matthew J. Wakefield, Evelyn Trounson, Douglas J. Hilton, Marnie E. Blewitt, Alicia Oshlack, and Ian J. Majewski. Chip-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 39(17):7415–7427, Sep 2011.

[184] Davide Gabellini, Michael R. Green, and Rossella Tupler. Inappropriate gene activation in fshd: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. *Cell*, 110(3):339–348, Aug 2002.

[185] S. Germain, J. Philippe, S. Fuchs, A. Lengronne, P. Corvol, and F. Pinet. Regulation of human renin secretion and gene transcription in calu-6 cells. *FEBS letters*, 407(2):177–183, Apr 1997.

[186] Glenn A. Maston, Sara K. Evans, and Michael R. Green. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics*, 7:29–59, 2006.

[187] Baoxu Pang and Michael P. Snyder. Systematic identification of silencers in human cells. *Nature genetics*, 52(3):254–263, Mar 2020.

[188] Heyuan Qi, Mingdong Liu, David W. Emery, and George Stamatoyannopoulos. Functional validation of a constitutive autonomous silencer element. *PloS one*, 10(4):e0124588, Apr 2015.

[189] Leah H. Matzat, Ryan K. Dale, Nellie Moshkovich, and Elissa P. Lei. Tissue-specific regulation of chromatin insulator function. *PLoS genetics*, 8(11):e1003069, Nov 2012.

[190] Kristen M. Kwan, Esther Fujimoto, Clemens Grabher, Benjamin D. Mangum, Melissa E. Hardy, Douglas S. Campbell, John M. Parant, H. Joseph Yost, John P. Kanki, and Chi-Bin Chien. The tol2kit: a multisite gateway-based construction kit for tol2 transposon transgenesis constructs. *Developmental dynamics: an official publication of the American Association of Anatomists*, 236(11):3088–3099, Nov 2007.

[191] Yangbin Gao and Yunde Zhao. Self-processing of ribozyme-flanked rnas into guide rnas in vitro and in vivo for crispr-mediated genome editing. *Journal of integrative plant biology*, 56(4):343–349, Apr 2014.

[192] Federico Katzen. Gateway(®) recombinational cloning: a biological operating system. *Expert opinion on drug discovery*, 2(4):571–589, Apr 2007.

[193] Adam G. West, Miklos Gaszner, and Gary Felsenfeld. Insulators: many functions, many mechanisms. *Genes development*, 16(3):271–288, Feb 2002.

[194] Santina Acuto, Rosalba Di Marzo, Roberta Calzolari, Elena Baiamonte, Aurelio Maggio, and Giovanni Spinelli. Functional characterization of the sea urchin sns chromatin insulator in erythroid cells. *Blood cells, molecules diseases*, 35(3):339–344, Nov 2005.

[195] J. H. Chung, M. Whiteley, and G. Felsenfeld. A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in drosophila. *Cell*, 74(3):505–514, Aug 1993.

[196] Ram P. Kumar, Jaya Krishnan, Narendra Pratap Singh, Lalji Singh, and Rakesh K. Mishra. Gata simple sequence repeats function as enhancer blocker boundaries. *Nature communications*, 4:1844, 2013.

[197] Jeroen Bussmann and Stefan Schulte-Merker. Rapid bac selection for tol2-mediated transgenesis in zebrafish. *Development*, 138(19):4327–4332, Oct 2011.

[198] Bonnie Burgess-Beusse, Catherine Farrell, Miklos Gaszner, Michael Litt, Vesco Mutskov, Felix Recillas-Targa, Melanie Simpson, Adam West, and Gary Felsenfeld. The insulation of genes from external enhancers and silencing chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 4:16433–16437, Dec 2002.

[199] Emery H. Bresnick, Hsiang-Ying Lee, Tohru Fujiwara, Kirby D. Johnson, and Sunduz Keles. Gata switches as developmental drivers. *The Journal of biological chemistry*, 285(41):31087–31093, Oct 2010.

[200] Jeffrey A. Grass, Meghan E. Boyer, Saumen Pal, Jing Wu, Mitchell J. Weiss, and Emery H. Bresnick. Gata-1-dependent transcriptional repression of gata-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15):8811–8816, Jul 2003.

[201] Frank Rosenbauer, Katharina Wagner, Jeffery L. Kutok, Hiromi Iwasaki, Michelle M. Le Beau, Yutaka Okuno, Koichi Akashi, Steven Fiering, and Daniel G. Tenen. Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, pu.1. *Nature genetics*, 36(6):624–630, Jun 2004.

[202] H. Dombret, M. P. Font, and F. Sigaux. A dominant transcriptional silencer located 5' to the human t-cell receptor v beta 2.2 gene segment which is activated in cell lines of thymic phenotype. *Nucleic acids research*, 24(14):2782–2789, Jul 1996.

[203] Amy T. Hark, Christopher J. Schoenherr, David J. Katz, Robert S. Ingram, John M. Levorse, and Shirley M. Tilghman. Ctcf mediates methylation-sensitive enhancer-blocking activity at the h19/igf2 locus. *Nature*, 405(6785):486–489, 2000.

[204] J. E. F. Butler. Enhancer-promoter specificity mediated by dpe or tata core promoter motifs. *Genes Development*, 15(19):2515–2519, 2001.

[205] B. D. Ortiz, D. Cado, V. Chen, P. W. Diaz, and A. Winoto. Adjacent dna elements dominantly restrict the ubiquitous activity of a novel chromatin-opening region to specific tissues. *The EMBO journal*, 16(16):5037–5045, Aug 1997.

[206] S. T. Smale. Core promoters: active contributors to combinatorial gene regulation. *Genes Development*, 15(19):2503–2508, 2001.

[207] W. Dietrich-Goetz, I. M. Kennedy, B. Levins, M. A. Stanley, and J. B. Clements. A cellular 65-kda protein recognizes the negative regulatory element of human papillomavirus late mrna. *Proceedings of the National Academy of Sciences of the United States of America*, 94(1):163–168, Jan 1997.

[208] P. A. Furth, L. St Onge, H. Böger, P. Gruss, M. Gossen, A. Kistner, H. Bujard, and L. Hennighausen. Temporal control of gene expression in transgenic mice by a tetracycline-responsive promoter. *Proceedings of the National Academy of Sciences of the United States of America*, 91(20):9302–9306, Sep 1994.

[209] M. Dunaway and P. Dröge. Transactivation of the xenopus rrna gene promoter by its enhancer. *Nature*, 341(6243):657–659, Oct 1989.

[210] T. Oelgeschläger, C. M. Chiang, and R. G. Roeder. Topology and reorganization of a human tfiid-promoter complex. *Nature*, 382(6593):735–738, Aug 1996.

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

[212] R. Korn, M. Schoor, H. Neuhaus, U. Henseling, R. Soininen, J. Zachgo, and A. Gossler. Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimaeric embryos with a high frequency and detect endogenous genes. *Mechanisms of development*, 39(1-2):95–109, Nov 1992.

[213] Serguei Parinov, Igor Kondrichin, Vladimir Korzh, and Alexander Emelyanov. Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo. *Developmental dynamics: an official publication of the American Association of Anatomists*, 231(2):449–459, Oct 2004.

[214] Dan Shen, Songlei Xue, Shuheng Chan, Yatong Sang, Saisai Wang, Yali Wang, Cai Chen, Bo Gao, Ferenc Mueller, and Chengyi Song. Enhancer trapping and annotation in zebrafish mediated with sleeping beauty, piggybac and tol2 transposons. *Genes*, 9(12), Dec 2018.

[215] Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, Dec 2014.

[216] Hyun Sik Jang, Woo Jung Shin, Jeong Eon Lee, and Jeong Tae Do. Cpg and non-cpg methylation in epigenetic gene regulation and brain function. *Genes*, 8(6), May 2017.