Computational Methods to Identify Regulatory Variants in the Non-coding Regions of the Human Genome

by

Shengcheng Dong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2021

Doctoral Committee:

Associate Professor Alan Boyle, Chair
Associate Professor Yuanfang Guan
Assistant Professor Jie Liu
Associate Professor Ryan Mills
Associate Professor Elizabeth Speliotes
Associate Professor Cristen Willer

Shengcheng Dong
shengchd@umich.edu
ORCID iD: 0000-0001-5728-8090

To my loving family

ACKNOWLEDGEMENTS

The work presented in this dissertation could not have been accomplished without the help of many people. First, I would like to thank my mentor Alan Boyle for the guidance and support. I did my first rotation in his lab as a great start here in Michigan. As an international student, I was nervous about the language and culture differences in my early years here, but the interaction with him always encouraged me. Thank you for making me feel welcome and supported throughout the 5 years. I am grateful for the opportunities you gave me to lead on my own projects, and your guidance from detailed technical problems to broad field overviews. The lab environment you created and your mentoring style made my study here a fun journey.

I would also like to thank all members of the Boyle lab. It has been a pleasure to work with each of you. Your enthusiasm for science and life always cheers my days in the lab. Adam Diehl has helped me with my projects since my rotation. Jessica Switzenberg gave me great help in writing manuscripts. I have enjoyed having peers in similar years working on the relevant projects. Thank you, Christopher Castro, Sam Zhao, and Ningxin Ouyang, for constructive discussions and common complaints on PhD life. I also learned a lot from the presentations from wet-lab students. Thank you all for making the Boyle lab a fun place to work.

In addition, I would like to thank my committee members. I enjoyed the discus-

sions during my committee meetings, and the suggestions and guidance you gave me were all critical to the accomplishment of my dissertation. I would like to thank Dr. Elizabeth Speliotes and Dr. Cristen Willer for their insights on GWAS studies, Dr. Jie Liu for suggestions on 3D conformation studies, Dr. Ryan Mills for advice on bioinformatics algorithms and coming up with the name for the tool I developed, and Dr. Yuanfang Guan for help on machine learning techniques.

I have also been fortunate to receive support from our collaborators. I would like to thank Yuanhai Luo and Ben Hitz for their help in building RegulomeDB. I have learned a lot about data curation from working with them.

I would like to thank Dr. Margit Burmeister, as my graduate advisor, for giving me lots of advice on my classwork and career development. I would also like to thank Julia Eussen for all the help through my PhD studies.

Finally, I would like to thank my family. To my fiancé, Mengyang Zhang, thank you for surviving our 5-year long-distance relationship. Thank you for taking the 2-hour flight over 60 times to see me throughout the years while also struggling with your PhD life. I would also like to thank my fiancé's parents for all their supports. To my parents, thank you for always supporting my choices in my life and encouraging me to study abroad. Thank you for everything you have done.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Evidence from Genome Wide Association Studies (GWAS) has provided us with insights into human phenotypes by identifying genetic variation statistically associated with diseases and complex traits. However, the functional consequences of these genetic variants remain unknown in many cases, especially for those in the non-coding regions of the human genome.

My dissertation focuses on single nucleotide polymorphisms (SNPs) as the most common genetic variation type. I define some SNPs as regulatory SNPs that can alter the transcription factor binding affinities within the DNA sequences of regulatory elements. This change affects downstream gene expression and plays a role in disease progression and trait development. Characterizing genome-wide regulatory variants is particularly challenging because the gene regulatory network is dynamic across various cell types and environmental conditions. In addition to the DNA sequence context, the gene regulatory network relies on epigenetic factors, such as chromatin accessibility, histone modification, and chromatin looping.

In this dissertation, I applied computational approaches to predict regulatory variants by incorporating sequence information and functional genomics annotations from various high-throughput assays. In chapter 2, I developed a computation tool, SURF, to prioritize the regulatory variants within promoters and enhancers with clinical relevance. These variants were validated by massively parallel reporter assays and used as an unbiased test set in CAGI5 "Regulation Saturation" challenge.

My algorithm achieved the best performance in this challenge compared to other participant groups.

In chapter 3, I extended SURF to TURF, a computational tool to predict tissue-specific functions of regulatory variants and provide a more robust prediction on genome-wide non-coding regions. By leveraging tissue-specific genomic annotations of tissues from the same organ, I also calculated TURF organ-specific scores covering most ENCODE project organs. Many of the GWAS traits showed enrichment of regulatory variants prioritized by TURF scores in their relevant organs, which indicates that these regulatory variants are likely to be involved in the trait developments and can be a valuable source for future studies.

In chapter 4, to enable the quick annotation on non-coding variants for the scientific community, I designed some major updates to an online tool, RegulomeDB. With the user's input of query variant, RegulomeDB returns the evidence from diverse functional genomics assays that overlaps the variant's position, displayed with interactive charts and a genome browser view. The new probabilistic score derived from SURF was also integrated into the query system. To further provide functional hypotheses to putative regulatory variants, I finally explored the pipeline to assign their target genes with evidence from eQTL studies and Hi-C experiments.

Together, my dissertation developed computational tools for broad community use on prioritizing and assigning target genes to regulatory variants in non-coding regions of the human genome.

CHAPTER I

Introduction

Characterizing the functional consequences of variants in the non-coding regions is a challenge in human genetics. In this dissertation, I applied computational methods to address this challenge. I first developed a computational tool to predict regulatory variants with training data from massively parallel reporter assays, and I extended it to predict tissue-specific function. I also designed a user-friendly online tool to enhance the use for the scientific community. Finally, I explored computational pipelines to assign target genes of the putative regulatory variants.

In this chapter, I will first describe the biological mechanism of regulatory elements and regulatory variants. I will then discuss the high-throughput functional genomics assays and the statistical analyses to map genome-wide regulatory elements and regulatory variants. Next, I will summarize the strategies and limitations of current computational tools to predict regulatory variants. Finally, I will discuss the computational approaches to assign the target genes to putative regulatory variants by 3D conformation assays.

1.1 Cis-regulatory elements and regulatory variants

This dissertation focuses on the cis-regulatory elements in the non-coding regions of human genome. These regulatory elements can be defined as enhancers, promot-

ers, insulators, and silencers based on their effects on gene expression level. The transcription factors (TFs) each recognize short DNA sequences (i.e., motifs) to bind with the regulatory elements and regulate their target genes expression [1]. The functional linkage between various TFs and their target regulatory elements is called 'gene regulatory network'. A growing amount of evidence shows that the dynamics of gene regulatory networks contribute significantly to diverse biological processes, including development, differentiation, and disease progression, emphasizing the importance of systematically characterizing the regulatory elements [1]. A large part of the challenge of deciphering gene regulatory networks comes from the fact that it not only relies on the DNA sequence (TF motif) itself but also depends on the chromatin factors such as chromatin accessibility, histone modification, and chromatin looping. These epigenetic factors vary a lot among cells and individuals. I will discuss the assays for genome-wide scanning on those chromatin factors in the following section.

In addition to characterizing regulatory elements, it is even harder to interpret the functional consequences of genetic variation within the regulatory elements. My dissertation focuses on predicting the functional consequences of the single nucleotide polymorphisms (SNPs) in regulatory elements, which is the most common type of genetic variation. SNPs can alter the TF binding sites, thus changing the binding affinity and further affecting downstream gene expression regulation, which are referred to as regulatory variants (or more specifically regulatory SNPs) (Figure 1.1). Numerous examples of regulatory SNPs are found to be associated with disease susceptibility, and some are validated to play an essential role in disease progression. For example, an SNP at the 1p13 cholesterol locus creates a TF binding site on minor alleles and alters the SORT1 gene expression level in liver, eventually increasing the risk for myocardial infarction [2]. As discussed before, predicting regulatory

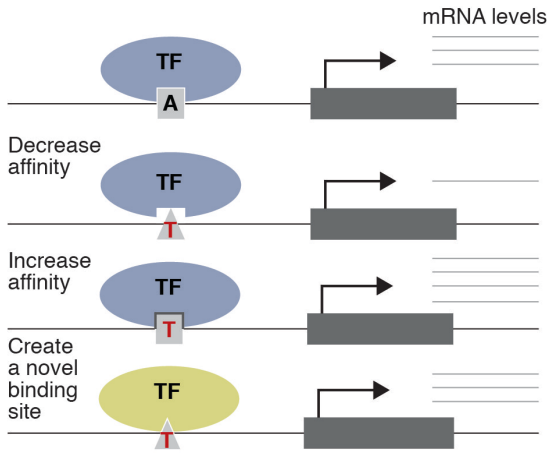


Figure 1.1: Schematic of potential mechanisms for a regulatory single nucleotide polymorphism (SNP) exerting effects on downstream gene expression. The A allele binds with a transcription factor (TF), which is an activator in this case. A genotype change from A to T will possibly 1) decrease TF binding affinity and downregulate gene expression; 2) increase TF binding affinity and upregulate gene expression; 3) create a novel binding site with another TF that affects other downstream gene expression.

variants involves incorporating regulatory sequence information with the chromatin factors. Furthermore, due to the dynamic of the gene regulatory network, evaluating the chromatin factors in cell type/tissue-specific context will provide more accurate prediction in studying diseases or traits relevant to particular cell types/tissues.

1.2 Functional genomics assays to identify genome-wide regulatory elements

With the extensive application of next-generation sequencing, various functional genomics assays are available to characterize genome-wide regulatory elements from different aspects (Figure 1.2).

1.2.1 DNase I hypersensitive sites sequencing (DNase-seq) and DNase footprints

DNase I hypersensitive sites (DHS) are the genomic regions that show hypersensitivity to the cleavage by DNase I endonucleases [4], thus representing a more open and loose structure of chromatin. DNase I hypersensitive sites sequencing (DNase-seq) is a technique to identify DHS on the whole genome, which combines traditional

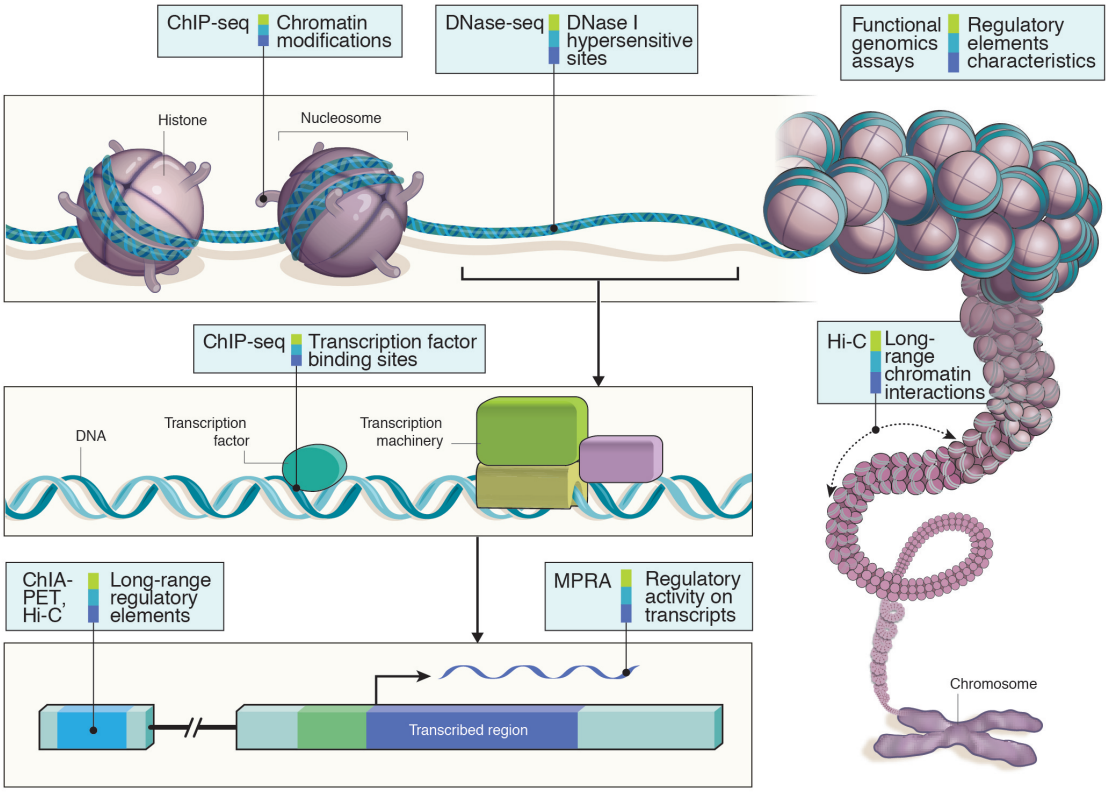


Figure 1.2: Examples of high-throughput functional genomics assays characterizing genome-wide regulatory elements from diverse aspects. Figure adapted from Ecker et al. 2012 [3].

DNase I footprinting and next-generation sequencing [5, 6]. The DHS are also interpreted as open chromatin regions accessible to proteins, including transcription factors (TF) to regulate gene transcription. DNase-seq has been a powerful assay to map genome-wide open chromatin regions among different cell types and treatment condition. ATAC-seq, as a more recently developed assay, offers an alternative way to map open chromatin regions based on the hyperactive Tn5 transposase [7]. ATAC-seq performs similar to DNase-seq in terms of sensitivity and specificity but requires fewer starting cell numbers and fewer preparation steps.

DNase footprints are a short fragment within a DHS that TFs bind that is protected by the digestion of DNase I. Traditionally, the DNase footprints were identified by running on a polyacrylamide gel [8].They now can be located on a genome-wide

scale by mapping the 'dips' in the middle of DNase-seq data signals. Several computational tools are available for this purpose. Some tools use the sequence information from TF motifs to assign the binding TFs for each DNase footprint [9, 10, 11, 12], while other tools only map the broad set of TF binding sites [13, 14, 15].

1.2.2 Chromatin Immunoprecipitation sequencing (ChIP-seq)

In contrast to DNase-seq mapping of open chromatin regions, Chromatin Immunoprecipitation sequencing (ChIP-seq) can pinpoint the binding sites of a specific TF. ChIP-seq relies on crosslinking proteins to DNA first, following by chromatin fragmentation. An antibody specific to a protein, such as a TF, is used in the immunoprecipitation step to bind and isolate the protein. Finally, the DNA fragments bound with the protein on the whole-genome are identified with next-generation sequencing techniques. In addition to TF, ChIP-seq can also identify the binding sites of histone modifiers, where a second crosslinker is sometimes needed since they bind more dynamically to chromatin [16].

The main limitation for ChIP-seq is the reliance on antibodies as a large portion of TFs do not have efficient antibodies. Tagging the protein of interest is a way to compensate for this limitation. However, careful design is needed to avoid the interruption of expression [17]. Meanwhile, the variation of antibody quality can lead to batch effects that require detailed control experiments. Furthermore, the application of ChIP-seq on cell types with low cell numbers is still challenging, especially for a transcription factor that does not have abundant binding sites on the genome.

Computational approaches have been applied to extract regulatory patterns from regions identified by ChIP-seq assays. I will introduce the two applications relevant to my dissertation here:

Position weight matrix (PWM)

A position weight matrix describes a TF binding preference, referred to as TF motif, by showing the frequencies of each base (A, C, G or T) on each position of the TF binding sequences [18]. The information content calculated from PWM represents the deviation from a uniform distribution of the four bases on each TF motif position. In some early studies, the TF binding sites were mostly identified from SELEX (systematic evolution of ligands by exponential enrichment) data [19]. More recently, PWMs are also computed by the TF binding sites from ChIP-seq data. Several databases collect these PWMs from various sources. For example, JASPAR curates a non-redundant set of PWMs from over 700 human TFs [20].

Despite the wide use of PWMs, one caveat is that the binding frequencies measured from PWMs might not be sufficient to predict the underlying binding affinities. Therefore, predicting TF binding sites by mapping PWM to the genome could potentially miss some true binding sites in underrepresented cell types. Moreover, the effects of variants on TF binding affinity could not be fully represented by the information content. Some computational tools were developed to overcome such limitations [21].

Chromatin state annotations

The combinations of chromatin modifications are found to be associated with certain regulatory elements, such as H3K4me3 associated with promoters, H3K4me1 and H3K27me3 associated with enhancers, and H3K9me3 associated with heterochromatin regions [22]. The chromatin modification patterns relating to various regulatory elements have been learned by applying computational approaches to histone mark ChIP-seq data. One of the most widely-used tools is chromHMM, which is based on a multivariate hidden Markov model. The primary chromHMM model infered 15 chromatin states from the observed patterns on 7 histone marks and named

these chromatin states based on different gene transcript activities, such as enhancer, active transcription start site and quiescent/low state. The chromatin state annotations from chromHMM are available for the 127 tissues in the Roadmap Epigenomics Project. Segway is another example that models histone mark patterns with gaussian mixture models [23, 24]. Noticeably, the chromatin states inferred from such tools are defined based on the prior biological knowledge of histone marks, which could contain some incorrect assignment for certain genomic regions.

1.2.3 Massively parallel reporter assay (MPRA)

In addition to techniques directly sequencing the regulatory elements, massively parallel reporter assay (MPRA) provides another functional measurement by assessing the effect on transcriptional activities from regulatory elements such as enhancers and promoters. MPRA relies on a library of plasmids that link all candidate regulatory sequences with a unique barcode and a reporter gene (GFP, LaZ, or others). The reporter vectors are then introduced into the cell lines or tissues of interest. The functional consequences of regulatory sequences can be assessed by counting the number of RNAs versus DNAs with their corresponding barcodes. Moreover, with saturation mutagenesis techniques, thousands of sequences with point mutations or small indels on regulatory elements can be synthesized in parallel with labeling by barcodes. Thus, MPRA can identify the functional consequences of thousands of mutant variants within the regulatory elements at one experiment.

MPRA was first applied by Patwardhan et al. in 2009 for studies on three bacteriophage promoters in vitro. More studies on genome-wide regulatory sequences were performed later in HepG2, GM12878, and K562 cell lines [25, 26]. Meanwhile, experiments on a limited number of promoters and enhancers, but with every possible point mutation on tested regulatory sequences were also conducted [27]. In detail,

Kircher and collaborators assessed 17,500 single nucleotide variants in 9 promoters and 5 enhancers with clinical relevance [28, 29, 26]. In addition to the limitation on tested cell types and genomic regions, another caveat for MPRA is that as a plasmid-based assay, the query sequences are removed from their genomic context. Therefore, the effects of chromosomal environment and enhancer-promoter looping is ignored in the MPRA experiments.

1.2.4 3D conformation assays

The regulatory elements can exert regulatory effects on their target genes across distances of kilobases, or even megabases in some cases, which rely on the spatial folding of chromosomes and chromatin looping [30, 31]. To map the three-dimensional conformation of chromosomes, genomic assays, including 3D conformation capture (3C), 3C-based technologies (4C,5C, and Hi-C), and chromatin-interaction analysis by paired-end-tag sequencing (ChIA-PET) have been developed [32, 33, 34, 35]. These assays have revealed chromosome 3D structures, such as topologically associated domains (TADs) with A/B compartments [36], and chromatin loops including enhancer-promoter interactions in a variety of human cells [34]. The experimental and computational approaches to identify those 3D structures will be discussed in more detail in this chapter's final section.

Despite the progress in developing various 3D conformation assays, the lack of benchmarks for assay performance has made it challenging to compare observations from different assays. The 4D nucleome project was launched to overcome this problem [37]. Moreover, efforts are being made to increase the resolution while reducing experimental noises by extending the Hi-C technology [38]. Meanwhile, single-cell Hi-C is another direction to advance the understanding of the heterogeneity of 3D conformation in cells, especially for cancer cells [39].

1.2.5 Functional genomics data from large consortia

While each of the functional genomics assays described previously characterized regulatory element landscape on various aspects, integrating the results from those assays will further enhance the understanding of regulatory elements. Large consortia have produced a vast number of datasets from diverse assays on human tissues or cell lines. Typical examples include the Encyclopedia of DNA elements (ENOCDE) project [40] and the Roadmap Epigenomics project [41]. The Roadmap project focused on samples taken directly from human tissues and cells. In contrast, the ENCODE project initially focused on human cells grown in culture and recently broadened coverage to primary human tissues and cells [40]. Up to now, over 6,000 ChIP-seq and DNase-seq datasets in human cells and tissues are available through the ENCODE data portal. Meanwhile, the 4D Nucleome (4DN) aims to develop experimental and computational approaches to measure genome conformation [37]. Genome-wide chromatin interaction maps from assays including Hi-C and ChIA-PET are available through the 4DN data portal.

One benefit from those large consortia is that uniform processing pipelines are being used. Therefore, we can compare the TF binding profiles and actively transcribed genes in diverse cell types and treatment conditions, which has significantly broadened our understanding of the gene regulatory networks in biological processes, such as differentiation, development and disease progression.

1.3 Statistical analyses to map genome-wide regulatory variants

Whole-genome sequencing has enabled genotyping genome-wide variants in large populations. Statistic models are widely applied in associating the genetic variation with variation from other high-throughput sequencing data or disease phenotypes to

map regulatory variants. Some typical applications are described in this section.

1.3.1 Quantitative trait loci (QTLs)

Quantitative trait loci (QTLs) are genome regions where genetic variation is statistically associated with variation in a quantitative trait [42]. The correlation is modeled by regressing the quantitative trait on variant genotypes, usually assuming additive allele effects. The most extensively conducted analysis is eQTLs, where the variants associated with gene expression levels from RNA-seq are mapped on the whole genome. However, identifying the causal variants from eQTLs is challenging because of the genetic correlation among variants, which is known as linkage disequilibrium (LD). Moreover, the presence of multiple causal variants within a locus requires a careful design in multiple testing correction. Several fine-mapping approaches are available to overcome this challenge [43, 44, 45], but they may still fail to detect the true causal variant when it is a rare variant or in high LD with many non-causal variants. Nonetheless, eQTL studies have increased our knowledge about the gene regulatory networks in diverse cell types and tissues. The largest eQTL datasets of human tissues come from the GTEx (The Genotype-Tissue Expression) project [46, 47]. To date, the GTEx project characterized genetic associations in 838 individuals over 49 human tissues, where a total of over four million variants are discovered associated with 20,000 genes. These eQTLs provide an invaluable resource for comparing gene regulatory networks across various tissues and assigning target genes to tissue-specific regulatory variants.

In addition to gene expression level, the quantitative trait can also be chromatin accessibility from DNase-seq or ATAC-seq, known as dsQTLs or caQTLs. dsQTLs were first identified in 70 Yoruba lymphoblastoid cell lines (LCL), which are shown enriched with TF binding sites [10]. More recently, caQTLs were called in 100 LCL

British samples [48] and more broadly in 1000 individuals from 10 diverse human populations [49]. As ATAC-seq being continuously generated, identifying caQTLs from more cell types or populations will be interested.

1.3.2 Allelic analysis on sequencing reads from ChIP-seq and RNA-seq

While QTLs require several samples to discover the association in genetic variation, allelic analysis on the two chromosomes within an individual provides another approach to measure the association. In this case, the quantitative trait can be the sequencing reads from TF ChIP-seq or RNA-seq around heterozygous sites within an individual. The allele-specific TF binding (ASB) variants or allele-specific gene expression (ASE) variants can be identified, which are the putative regulatory variants showing variation in TF binding affinity or gene expression. The main advantage of such an approach is the natural controls within an individual, thus avoiding the normalization step when comparing different individuals. However, identifying such regulatory variants is restricted by the number of available heterozygous SNPs in each individual. Also, the computational cost is relatively high in the alignment process to avoid mapping bias to the maternal and paternal genome since the allelic analysis with a limited number of reads is more sensitive to such bias. Previous studies have identified thousands of ASB and ASE variants, mainly in LCL samples [50, 51, 52, 53, 54, 55]. Noticeably, some tools further leverage the allelic imbalance from allelic analysis on RNA-seq reads to refine mapping for eQTL studies [56].

1.3.3 Genome-wide association studies (GWAS)

Another example of the quantitative trait is the disease risk or trait. Genome-wide association studies (GWAS) have discovered thousands to millions of genetic variations associated with diseases and traits, including schizophrenia [57], inflamma-

tory bowel disease [58], body mass index [59], and many others. A massive number of novel risk loci of those diseases and traits have been identified successfully, leading to follow-up studies further explaining the underlying biological mechanisms. Moreover, GWAS results can guide identifying the individuals at high risk of certain diseases, thus having wide application in precision medicine. GWAS Catalog data portal contains a curated collection of GWAS and their results, including more than 70,000 variant-trait associations up to now [60].

Despite the success in understanding certain diseases, many of the associations between variants and traits remain unexplained, especially for those ∼90% variants in the non-coding regions [61]. The major difficulty comes from the linkage disequilibrium among variants, as discussed previously for QTL studies. As a result, GWAS does not necessarily pinpoint the disease causal variant, and it is more challenging for complex traits involving a large number of genes. In such cases, the GWAS loci typically have small effect sizes, and the causal variants are even harder to be identified if they are rare variants. Moreover, the ethnic differences in disease risks are overlooked in many of the current GWAS results. These limitations can be improved with a larger sample size and more diverse populations.

1.4 Computational tools to predict regulatory variants

Many computational tools are available to predict regulatory variants in non-coding regions with different scoring schemes to leverage the evidence from functional genomics assays and association studies (Table 1.1). They are widely used to refine the functional regulatory variants from a list of candidates from association studies, such as GWAS.

Table 1.1: Commonly-used computational tools for prioritizing regulatory variants in non-coding regions

Name	Target of predictors	Algorithm models	Tissue-specific?	Interface	Reference
RegulomeDB	Regulatory function	Empirical decision tree	No	Website	(Boyle et al. 2012)
HaploReg	Regulatory function	Enrichment analysis	Yes	Website	(Ward and Kellis 2016)
DeepSEA	Regulatory function	Deep learning	No	Website & command line	(Zhou and Troyanskaya 2015)
GWAVA	Disease risk	Random forest	No	Website	(Ritchie et al. 2014)
GenoSkyline	Disease risk	Unsupervised learning	Yes	Precalculated scores	(Lu et al. 2016)
fitCons	Fitness consequence	Clustering	No	Precalculated scores	(Gulko et al. 2015)
LINSIGHT	Fitness consequence	Generalized linear model	No	Command line & Precalculated scores	(Y.-F. Huang, Gulko, and Siepel 2017)
FunSeq2	Regulatory function	Weighted scoring based on entropy	No	Website	(Fu et al. 2014)
FUN-LDA	Regulatory function	Latent Dirichlet Allocation model	Yes	Website	(Backenroth et al. 2018)
GenoNet	Regulatory function	Semi-supervised	Yes	Website	(He et al. 2018)
GWAS4D	Regulatory function	Jointly likelihood framework (Cepip from Li et al. 2017)	Yes	Website	(D. Huang et al. 2018)
gkm-SVM	Regulatory function	SVM	Yes (but in limited tissues)	Command line & Precalculated scores	(Beer 2017)
FATHMM-MKL	Disease risk	SVM	No	Website	(Shihab et al. 2015)

1.4.1 Overlapping functional genomics annotations with query variant position

The most straightforward strategy to annotate non-coding variants is intersecting the functional annotations of regulatory elements with the query variant position. The query variants can then be scored or prioritized with the emphasis on specific

evidence from empirical knowledge. For example, RegulomeDB 1.1 provides a ranking score for each query variant from a manually designed decision tree [62]. It emphasizes the evidence from eQTLs and DNase footprints along with other annotations, including open chromatin regions and sequences matching TF motifs. In addition, the details of each hit on functional annotation are available through a website interface. Another widely-used computational tool is HaploReg v4 [63], which incorporates functional genomics annotations with the map of haplotype blocks. In this way, the variants in linkage disequilibrium (LD) with the query variant including itself will be annotated. It is typically useful for query variants from GWAS as it is affected by LD blocks. HaploReg website also presents a summarized table of available evidence on query variants. One benefit of such a straightforward strategy is that the results are usually easy to interpret. However, extra filtering steps will need to be done manually to prioritize the candidate variants according to the annotation results.

1.4.2 Application of machine learning techniques

Other than simple intersections and empirical pipelines, machine learning techniques have been widely used. By incorporating functional genomics annotations in a more unbiased way, they provide more accurate and robust predictions on query variants with continuous scores. For example, supervised machine learning methods, including support vector machine (SVM) and random forest, are used in tools such as gkm-SVM [64], FATHMM-MKL [65], and GWAVA [66]. More recently, deep learning methods are also applied in many tools, which will be discussed in detail in the following section. Such supervised machine learning models require a prelabeled training set. The variation in the source of training set among tools leads to different targets of predictions. These predictions can be grouped into three main

categories: 1) Disease risk predictors; 2) Fitness consequence predictors; 3) Regulatory function predictors. Their typical sources of training set are 1) Pathogenic or disease-related variants from HGMD database [67] or GWAS studies; 2) Deleterious versus neutral variants from evolutionary studies; 3) Functional regulatory variants from eQTLs or reporter assays. Other tools applied unsupervised or semi-supervised methods to learn inherent patterns within the training data, which is less common [68, 69]. Compared to predictions on an organism level, only a few tools provided cell type/tissue-specific scores and mostly relied on the epigenetic data from the Roadmap Epigenomics Project in 127 tissues [68, 69, 70, 71].

1.4.3 Deep learning in predicting regulatory variants

Deep learning is a rapidly evolving field in prioritizing regulatory variants. Contrary to standard machine learning methods, deep learning models can automatically learn more complicated patterns with less handcrafting. Following the successful application in natural-language processing, deep learning has been applied in mining regulatory grammars from DNA sequences to predict the functional consequences of non-coding variants. For example, DeepSEA trained a convolutional neural network to predict 919 functional genomics features, including TF binding, open chromatin, and histone mark profiles, across various cell types in the training set [72]. Then the 919 predictors were adapted to predict variant effects on regulatory function. One significant benefit of this multitasking model is that the predictive features on sequence patterns can be learned jointly and shared when predicting diverse chromatin profiles. The same team has extended their work to learn features on gene expression levels and predict regulatory variants in autism spectrum disorder [73]. Meanwhile, deep learning was also applied in predicting pathogenic variants in human diseases from population sequencing [74] .

One major limitation of the deep learning method is that it requires a large training set. Also, careful curation on confounders and data bias is needed to avoid the overfitting problem. Despite high accuracy, it is often hard to interpret deep learning results since the essential features cannot be easily extracted compared to traditional machine learning methods. Some interpretation tools such as DeepLift [75] and LIME [76] overcome this problem to some extent, while more studies are still required.

1.4.4 Challenges for current computational tools

Due to the diverse source of training data, the existing computational tools tend to capture various characteristics of functional regulatory elements. As a result, the scores from those tools show inconsistent performance across test sets on pathogenic, neutral, or regulatory variants from a previous study [77]. These differences should be taken into account to choose appropriate tools. Moreover, the lack of a gold standard makes it hard to compare performance across different tools, even for those in the same category of target predictor. For predicting functional regulatory variants, which is the focus of my dissertation, it is common to use test variants from eQTLs, dsQTLs, and MPRA when evaluating performance. However, the QTLs are not necessarily the variants causing regulatory variation, and MPRA experiments are limited in the variation of tested cell types and genomic regions. A more thorough comparison is needed as the number of experimentally validated variants is increasing. Thus, the potential overfitting problem on well-studied cell types/tissues can then be more carefully examined. Also, predicting regulatory variants with cell type/tissue-specific function is a growing interest in specific disease or trait research.

On the other hand, it is also challenging to retrieve and integrate functional genomics annotation from numerous assays through large consortia, including EN-

CODE, Roadmap, and GTEx. Manual curations and computational pipelines are required to make maximum usage of those resources from diverse tissues and cell types. Moreover, a user-friendly interface is also preferred in addition to the algorithm itself to facilitate the research from a broad community.

1.5 Assigning target genes based on chromatin 3D structures

The spatial distance between the regulatory elements and downstream genes provides a more meaningful measurement than the traditional linear distance from chromosome coordinates to assign target genes for putative regulatory variants. Technologies based on Chromosome Conformation Capture (3C), such as Hi-C, have significantly advanced our understanding of chromatin 3D structures. In Hi-C experiments, chromatin is crosslinked, then digested, and re-ligated so that the DNA fragments physically close in the 3D organization of the genome will form ligation products [78]. The Hi-C heat map represents the interaction frequency of genomic regions, usually in 10kb to 1Mb bins. I will discuss two chromatin structures from Hi-C heat maps (Figure 1.3).

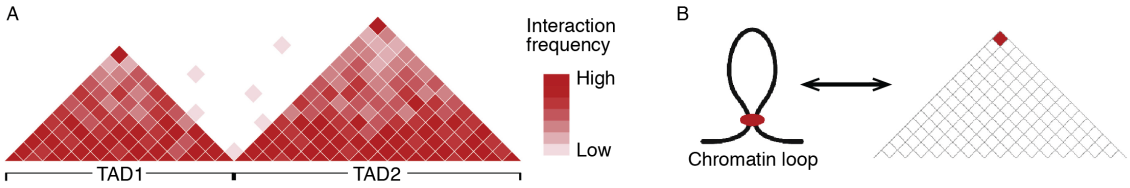


Figure 1.3: Schematic of topologically associating domains (TADs) and chromatin loop identified from Hi-C heat map. A) Two TADs with high intradomain interaction frequencies but low interdomain interaction frequencies. B) A simplified Hi-C heat map containing a chromatin loop structure, representing by a corner dot. Figure adapted from Chang, Ghosh, and Noordermeer 2020 [79].

1.5.1 Topologically associating domains (TADs)

Topologically associating domains (TADs) were discovered in Hi-C heat map [80] (Figure 1.3A), where the interaction frequencies between DNA sequences within domains are relatively higher than the frequencies outside domains. While initially defined in low resolution (40kb), technique improvements have been made to increase the resolution and discovered finer-scale structures, such as subTADs [79].

The biological function of TADs tends to be context-dependent and remains unclear in many cases. Nevertheless, it is believed that TADs serve as gene regulation units, enriched with coregulated gene clusters and enhancer-promoter pairs [79]. A large proportion of the boundaries of the TADs bind with the CTCF protein (CCCTC-binding factor). However, CTCF binding disruption does not impact a higher-order chromatin structure, which is called a compartment [81]. Two types of compartments were found from the low-resolution Hi-C heatmap: A compartment contains mainly actively transcribed genes while B compartment consists of gene-poor lamina-associated domains (LADs) [33]. More recently, higher-resolution Hi-C heat maps further partitioned the two compartments into at least six smaller subcompartments marked by various combinations of histone modifications [34].

1.5.2 Chromatin loops

The corner dot structures in the Hi-C heat map represent long-range looping interactions (Figure 1.3). The loop domains on the boundary of TADs are driven by CTCF and cohesion complex, which are responsible for establishing the TADs. While inside the TADs containing actively transcribed genes, a large fraction of the loop domains represents the strong contacts of enhancer-promoter pairs. These looping interactions are useful to assign target genes for regulatory elements. However, addi-

tional information such as histone modifications is needed to annotate the regulatory elements in the loop domains.

Meanwhile, promoter-capture Hi-C is developed to more specifically capture the interactions between enhancers and promoters in diverse cell types/tissues [82, 83]. Also, chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) can map long-range looping interactions with the enrichment step for chromatin complex containing a specific protein, for example the RNA Polymerase II from transcription-initiation complex to identify the enhancer-promoter pairs.

In summary, the TADs and chromatin loops identified from 3D conformation assays measure long-range genomic interactions, including the interactions between regulatory elements and target genes. Therefore, incorporating tissue-specific TADs and chromatin loops is useful for understanding the functions of regulatory variants. However, incorporating such datasets can be challenging since they are from various sources involving different cell types and processing steps.

CHAPTER II

Predicting Non-coding Variant Effects in Disease-Associated Promoters and Enhancers from MPRA Experiments

2.1 Abstract

Here we present a computational model, SURF (Score of Unified Regulatory Features), that predicts functional variants in enhancer and promoter elements. SURF is trained on data from massively parallel reporter assays and predicts the effect of variants on reporter expression levels. It achieved the top performance in the Fifth Critical Assessment of Genome Interpretation "Regulation Saturation" challenge. We also show that features queried through RegulomeDB, which are direct annotations from functional genomics data, help improve prediction accuracy beyond transfer learning features from DNA sequence-based deep learning models. Some of the most important features include DNase footprints, especially when coupled with complementary ChIP-seq data. Furthermore, we found our model achieved good performance on predicting allele-specific transcription factor binding events. As an extension to the current scoring system in RegulomeDB, we expect our computational model to prioritize variants in regulatory regions, thus help the understanding of functional variants in noncoding regions that lead to disease.

2.2 Introduction

Evidence from Genome Wide Association Studies (GWAS) has provided us with insights into human phenotypes by identifying variation statistically associated with diseases [84]. However, GWAS is confounded by linkage disequilibrium when identifying the causal variants. Thus, it is desirable to extend these studies beyond association to an understanding of biological impact. Unfortunately, determining the function of these variants remains a major challenge, especially for single-nucleotide polymorphisms (SNPs) in non-coding regions of the genome, where most of these GWAS variants fall [85].

The advent of functional genomics assays has assisted us in mapping disease causative SNPs from GWAS. By intersecting the position of variants with regulatory elements identified from these assays, computational tools have been developed to prioritize SNPs in non-coding regions [86] . Tools such as RegulomeDB [62], GWAS3D [87], and HaploReg [63] have reduced time-consuming experiments for validation. Machine learning methods have been widely applied to integrate the annotations from functional genomics assays in a more sophisticated way, and thus produce more robust and accurate predictions [88]. More recently, the rapid development of deep learning techniques has enabled mining in high-dimensional sequences data. Some examples include DeepSEA [72], DeepBind [89], DanQ [90], Define [91], and Basenji [92]. However, since data sets used for training in those algorithms vary, comparisons across different models can become a problem considering there is currently no gold-standard for evaluation [86].

One independent method for evaluating the performance of these tools is through the use of massively parallel reporter assays (MPRA) wherein libraries that are de-

rived from PCR-based saturation mutagenesis have been applied to test the effect of variants in a putative regulatory region. These assays can measure the functional effect of variants on the expression level of a reporter construct in a high-throughput manner allowing for rapid testing of large numbers of variants. Kircher and collaborators performed MPRA for 17,500 single nucleotide variants (SNVs) in 9 promoters and 5 enhancers with clinical relevance [28, 29, 26]. This dataset allows for an unbiased comparison of computational tools used for variant prioritization and was used in this manner for the Fifth Critical Assessment of Genome Interpretation (CAGI5) "Regulation Saturation" challenge. Participants were asked to predict the functional effects of variants in these regulatory regions as measured by the reporter expression.

We present a machine learning-based computational framework, SURF (Score of Unified Regulatory Features), which combines features from RegulomeDB and DeepSEA, to predict the effect of variants on expression in promoters and enhancers. Our model achieved the top performance in the CAGI5 "Regulation Saturation" challenge. We also demonstrate that direct features from functional genomics data improve the prediction accuracy in addition to features from DNA sequence-based deep learning models.

2.3 Methods

2.3.1 Datasets in CAGI5 Regulation Saturation Challenge

The Regulation Saturation Challenge assessed 17,500 SNVs in 5 human disease associated enhancers (IRF4, IRF6, MYC, SORT1, ZFAND3) and 9 promoters (F9, GP1BB, HBB, HBG, HNF4A, LDLR, MSMB, PKLR, TERT) in a massively parallel reporter assay (Figure 2.1A). The MPRA libraries were derived from saturation mutagenesis of regulatory regions up to 600bp length, with a random change rate of 1 per 100 bases.

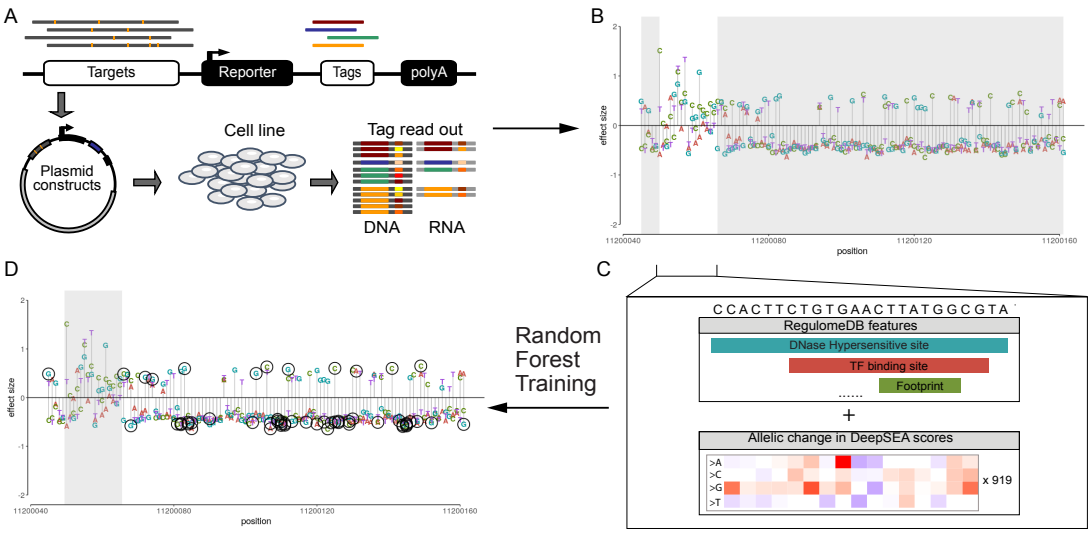


Figure 2.1: Workflow of SURF. A) The effect of variants in promoters and enhancers was tested through massively parallel reporter assays (MPRA). B) Effect size modeled from regression for each variant was provided with 25% of data (white area) used for training and 75% of data (grey area) hidden from participants. C) A multiclass random forest model is trained by combining features from RegulomeDB and DeepSEA on training data. D) Prediction of variants with significant effects (circled points) is made from random forest models.

Approximately 25% of all measured SNVs were used for training (4,650 SNVs in total), and the remaining 75% of the data were held from competitors and used for testing by an independent assessor. The count of transcribed RNA and DNA of the transfected plasmid library was modeled by applying multiple linear regression (Figure 2.1B). The coefficients ("effect size") and re-scaled p-values ("confidence score") from regression were provided in the training set. The SNV with a confidence scores greater or equal to 0.1 (i.e. p-value of 1×10^{-5}) was defined as "has an expression effect".

2.3.2 Tasks in CAGI5 Regulation Saturation Challenge

For each variant in testing set, the participants were asked to submit prediction of effect of the variant in one of the three cases: repressive, activating, or no

effect ("Direction"), and the probability of a correct assignment of the prediction ("P_Direction"). The participants also needed to submit a prediction of the confidence score for each variant, as well as the standard error of the prediction ("SD").

2.3.3 Model training

For each variant in training and test data, we created features from functional genomics data retrieved from RegulomeDB [62]. We also used sequence-based features from DeepSEA [72]. We further trained a random forest model to predict direction of variant effects and confidence score (Figure 2.1).

The first six features were created by querying each variant through RegulomeDB (Boyle et al. 2012). All ENCODE data represented in RegulomeDB is from the 2012 freeze and subsequent publication. We assigned binary values to represent if the position of the queried variant overlaps the following functional genomics regions:

1. Transcription factor (TF) binding site

TF ChIP-seq peaks were from ENCODE data.

2. Open chromatin site

DNase-peaks were from ENCODE data.

3. TF motifs

TF motif matches were called using positional weight matrices (PWM) from RegulomeDB. Positional weight matrices were from TRANSFAC [93], JASPAR CORE [94], UniPROBE [95] and Jolma et al [96].

4. Matched TF motif

TF motif matches were obtained as described in feature 3, but further requiring the PWM motif matching with a TF binding peak of the same TF from ChIP-seq in the same position.

5. DNase footprint

DNase footprints were called by combining PWMs and DNase-seq data sets. We used footprint calls from [97], Pique-Regi et al [98] and Piper et al [13].

6. Matched DNase footprint

DNase footprints were obtained as described in feature 5, but further requiring the PWM motif matching with a TF binding peak from ChIP-seq in the same position.

We also included additional numeric features:

7. ChIP-seq signal

We calculated the maximum TF ChIP-seq signal from feature 1 for each position in the regulatory regions.

8. Maximum information content change of TF motif

For each variant, we calculated the information content change of PWMs called in feature 3 and took the one with maximum absolute value.

9. Maximum information content change of matched TF motif

For each variant, we calculated the information content change of matched PWMs called in feature 4 and took the one with maximum absolute value.

10. DeepSEA scores

We passed a vcf file of all variants through DeepSEA model to predict chromatin effects of each mutation on 919 functional genomics features, including chromatin accessibility, TF binding and histone modification. We used the difference between reference and alternative alleles of those 919 functional genomics features in our model. We also included the functional significance score for each variant, which considers chromatin effects as well as evolutionary conservation.

A random forest model was trained to make predictions for both direction of effects and confidence scores. Specifically, we used the R package *randomForest* version 4.6-12 with ntree=500 [99]. For direction prediction, we first classified training data

from all studied regulatory regions into three groups using the following criteria:

1. Repressive (-1): confidence greater than or equal to 0.1 and effect size smaller than 0 (736 in total).

2. Activating (+1): confidence greater than or equal to 0.1 and effect size greater than 0 (374 in total).

3. No effect (0): confidence smaller than 0.1 (3,540 in total).

We then trained three binary classifiers for each label with a random forest model and predicted the label with the highest probability. We assigned "P_Direction" column with the prediction probability from the model. In order to generate a confidence prediction, we trained a random forest regression model on confidence scores and calculated the standard deviation of predictions from 500 trees in "SD" column.

2.3.4 Performance evaluation

Group performance was evaluated on correlation coefficients and the area under the receiver operating characteristic (AUROC). Pearson and Spearman correlation coefficients were calculated for predicted direction and effect size from MPRA on variants in test set in the same way as the assessors. Three categories of AUROC were assessed: variants with positive effects versus negative effects, variants with positive effects versus all variants, and variants with negative effects versus all variants. Predicted directions were treated as labels and effect sizes were used as probability scores. To increase the sensitivity of model comparisons, we also provided continuous value predictions as requested by the assessors, which are a transformation from

"P_Direction":

$$Direction' = \begin{cases} P\_Direction & \text{if Direction=1} \\ -P\_Direction & \text{if Direction=-1} \\ 1 - P\_Direction & \text{if Direction=0 and } D_{-1} < D_{+1} \\ P\_Direction - 1 & \text{if Direction=0 and } D_{-1} > D_{+1} \end{cases}$$

where D_i is the probability of class $i (i = -1, 0, +1)$ from random forest model.

Pearson correlation with continuous predictions were reevaluated among top three methods by the assessors (Table 2.2).

2.3.5 Allele-specific transcription factor (TF) binding analysis

Allele-specific TF binding sites were defined as variants that result in stronger binding of a TF to one allele at heterozygous sites in an individual. We applied AlleleDB pipeline to call allele-specific TF binding sites using ChIP-seq data downloaded from ENCODE project [52]. 1,814 allele-specific binding sites were called in GM12878 cell line from 76 TFs at an FDR of 5%. To test the performance of our binary classifier trained on CAGI5 data, we also built a control set including 10,783 variants having equal ChIP-seq read counts on two alleles at heterozygous sites. For all 48,630 heterozygous sites, we calculated the allelic ratio defined by the ratio between number of ChIP-seq reads from the allele with stronger binding affinity and total number of reads from two alleles. For cases where multiple TFs shared a heterozygous variant, we took the maximum ratio.

2.4 Results

2.4.1 SURF outperforms other groups in CAGI5 Regulation Saturation Challenge

SURF combines features from RegulomeDB, which directly intersects variants with functional genomics annotations, and DeepSEA, which generates transfer learn-

ing features from genomics assays. For assessment, both Pearson and Spearman correlation coefficients were calculated for predicted direction and effect size from MPRA on test data. To examine how false positive rate changes with true positive rate, the area under the receiver operating characteristic (AUROC) was also calculated (Table 2.1). Overall, we were close to group 7 on correlation coefficients, and we outperformed all groups in terms of all three categories of AUROC, especially in the case when distinguishing between variants with positive and those with negative effects on expression level. In addition, we note that it is generally easier to predict negative effects compared with positive effects, which might because there were more examples with negative effects in training set.

Table 2.1: Correlation and AUROC for predicting direction of variant effects across all participated groups. The best submission of each group was selected and the best performance of each category is bolded. AUPRC and correlation with continuous prediction scores were calculated.

Participant (lab-submission)	**Pearson correlation**	**Spearman correlation**	**Pos V Neg AUROC**	**Pos V Rest AUROC**	**Neg V Rest AUROC**
3-4 (our group)	0.301	0.239	**0.842**	**0.716**	**0.835**
7-3	**0.318**	**0.249**	0.762	0.706	0.776
5-6	0.255	0.235	0.714	0.608	0.691
1-2	0.069	0.046	0.544	0.553	0.636
6-1	0.103	0.094	0.537	0.544	0.584
4-2	0.041	0.033	0.556	0.528	0.571

2.4.2 Model performance in different enhancers and promoters

We assessed our performance in each of the 5 enhancers and 9 promoters (Figure 2.2). Continuous value predictions were used for calculating Pearson correlation with effect sizes. We observe no evident difference in performance between enhancers and promoters, but predictions on enhancers are more consistent in terms of AUROC performance. Also, our model performance has no strong association with cell types. The four regions in HEK293T (HNF4A, MSMB, TERT and MYC) have a wide range of performance. Overall, we predicted most accurately in regions of: MYC (HEK293T), PKLR (K562) and HBB (HEL_92.1.7). Interestingly, the cell line HEL_92.1.7 has no corresponding functional genomics data from the ENCODE project. In addition, ZFAND3 data is from mouse pancreatic beta cell lines (MIN6). These imply our model is able to predict these effects from the available data in other cell types.

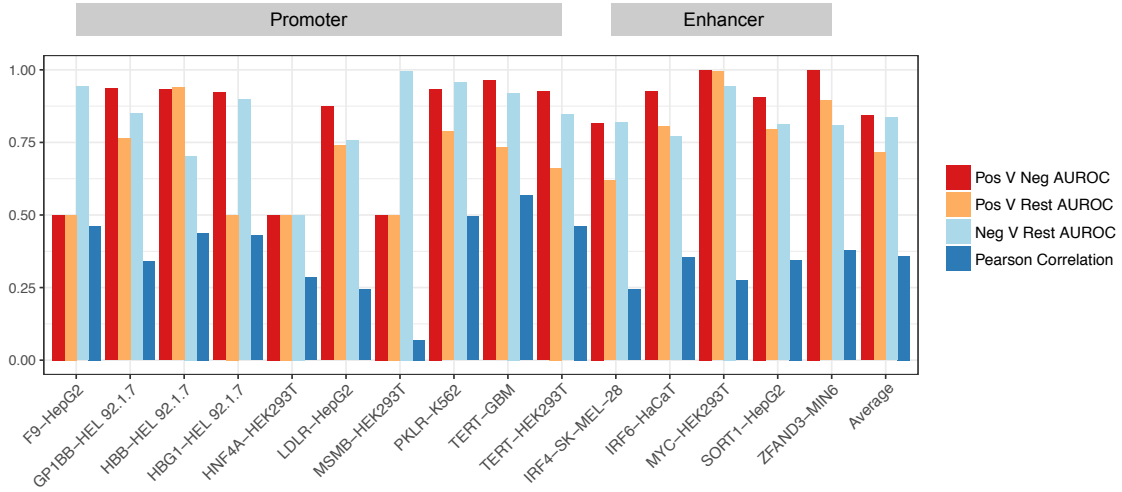


Figure 2.2: Performance across regions. Cell type names are appended at the end of promoter and enhancer regions. The average performance across all regions is also shown.

2.4.3 Features from RegulomeDB provide complementary information to DeepSEA scores

We next analyzed the predictive importance of RegulomeDB features. We calculated Pearson correlation of features and absolute value of effect sizes in test data (Figure 2.3A). All features have positive correlation, which is consistent with the fact that the variants in functional regulatory elements have a higher chance of affecting the expression level downstream. Among all binary features from RegulomeDB, features such as matched TF motif and matched DNase footprint have the highest correlation coefficients, which indicates that integrating sequence information with evidence from functional genomics data directly into one feature assists prediction accuracy. We further examined two of the most predictive features in the region of MYC enhancer, where we achieved the best AUROC compared with other enhancers and promoters. As shown in Figure 2.3B, these two features from RegulomeDB, DNase footprint and matched DNase footprint, are largely in agreement with the position of variants leading to significant change of gene expression beyond DeepSEA scores.

2.4.4 Predicting allele-specific TF binding events

To test the generality of our model, we next evaluated how SURF performs on predicting allele-specific TF binding events identified from ChIP-seq data. We collected 1,848 variants associated with allele-specific binding in GM12878 cell line, and then generated prediction scores using the binary classifier we trained from variants with no effects versus the rest of the variants in CAGI5 training set. Overall, our model is able to predict allele-specific binding events with a fairly good performance (AUROC=0.6218; AUPRC=0.2298). We further relaxed our thresholds to examine the performance on a wider spectrum of allelic ratio, which is defined by the ratio

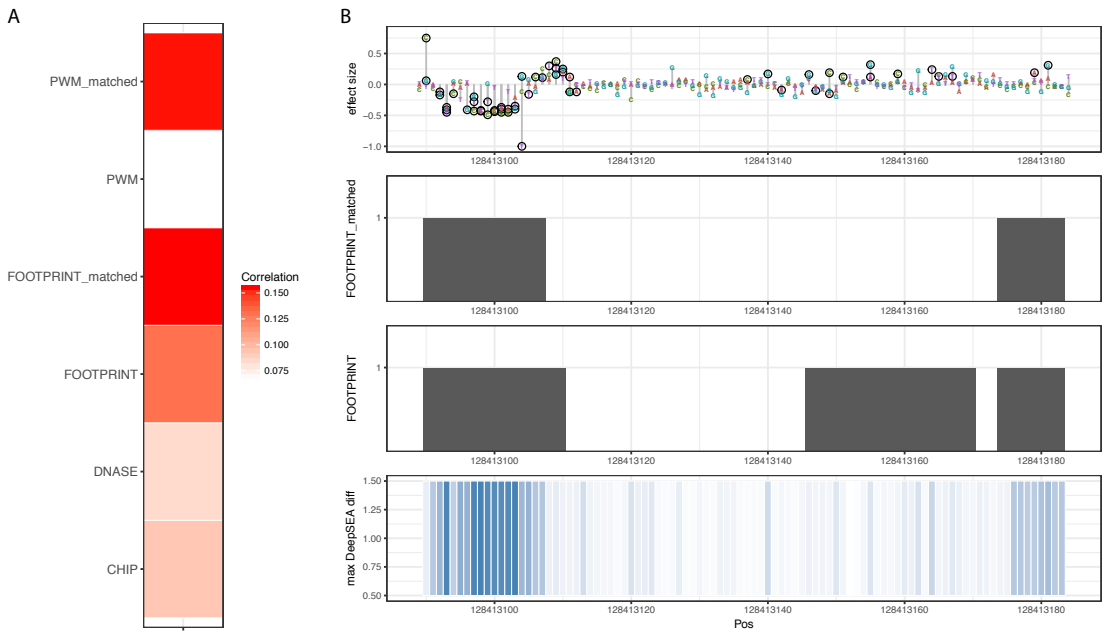


Figure 2.3: Features from RegulomeDB facilitate prediction. A) Pearson correlation of features from RegulomeDB and absolute value of effect sizes from MPRA in test data. B) A region of the MYC enhancer in HEK293T cell line showing measured MPRA data with SNVs having significant effect circled. Two binary features from RegulomeDB (DNase footprint and DNase footprint with matched TF ChIP-seq peak) show agreement with the position of these variants. DeepSEA scores also identify some of the functional variants in this enhancer.

between number of ChIP-seq reads from the allele with stronger binding affinity and total number of reads from two alleles. We found a significant difference in prediction scores for heterozygous sites showing balanced (allelic ratio smaller than 0.6) and imbalanced (allelic ratio equal or larger than 0.9) TF binding affinity (Figure 2.4, p-value = 9.735e-311 from a t-test).

2.5 Discussion

Understanding the function of variants in noncoding regions remains a major challenge to interpret results from GWAS studies. The CAGI5 Regulation Saturation challenge has provided a valuable dataset for developing prediction models on regulatory variants leading to significant effects on expression level. Here we described

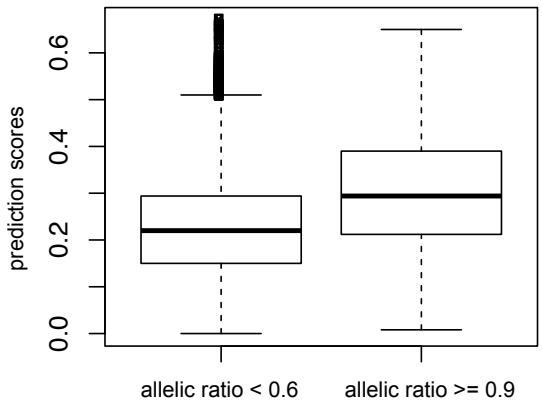


Figure 2.4: Boxplot of prediction scores for heterozygous sites showing balanced and imbalanced TF binding affinity from two alleles. Allelic ratio is calculated by the number of ChIP-seq reads from the allele with stronger binding affinity divided by total number of reads from two alleles.

our model, SURF, based on our existing resource RegulomeDB, that achieves the top performance in this challenge. However, one limitation of the evaluation with AUROC is that the imbalance rate was different across groups, which makes it hard to compare. A more accurate comparison is the correlation between continuous prediction scores and effect sizes from MPRA, which is shown in Table 2.2 but only available from three groups.

We found that the direct annotations from functional genomics data queried through RegulomeDB enables the improvement of prediction beyond the transfer learning features from the DeepSEA model. One possible reason to explain the improvement is that the chromatin features from underrepresented cell types in deep learning model are compensated by direct annotations from RegulomeDB. Thus, continued working on RegulomeDB resource, including updates and expansion of available data from ENCODE project, will enable us to develop prediction models with better accuracy. For example, 3D chromatin interaction data illustrating loops between enhancers and promoters can be used to assign target genes of variants in regulatory elements. In addition, ATAC-seq as an alternative method for study-

Table 2.2: AUPRC for predicting direction of variant effects across all participated groups and Pearson correlation with continuous scores. Only the first three groups have available scores.

Participant (lab-submission)	Pos V Neg AUPRC	Pos V Rest AUPRC	Neg V Rest AUPRC	Pearson correlation with continuous scores
3-4 (our group)	0.637	0.097	0.308	0.452
7-3	0.611	0.165	0.312	0.451
5-6	0.639	0.261	0.434	0.277
1-2	0.446	0.051	0.147	NA
6-1	0.007	0.004	0.680	NA
4-2	0.576	0.063	0.079	NA

ing chromatin accessibility will potentially give us complementary information to DNase-seq.

Furthermore, instead of obtaining general features through all available cell types in RegulomeDB as we did in this challenge, it is possible to query features in a cell type-specific way to improve performance. Although a previous study suggests that limiting features to be cell type specific does not increase prediction accuracy for MPRA data [100], it is worth exploring further whether this is due to the limitation of MPRA to capture cell type-specific activity. Another strategy is to integrate cell type-specific features with a generic model trained with all available cell types, thus taking advantage of a sufficient set of training data as well as a retention of cell type-specific information.

The initial premise behind the development and scoring in the RegulomeDB tool was that functional genomics data is key to understanding and prioritizing variants

that may be disrupting transcription factor binding and thus having a direct effect on gene expression. We have shown that these data have aided our model to perform well on MPRA training data and improve the ability to predict allele-specific TF binding events. Multiple studies have successfully applied RegulomeDB to infer regulatory variants in cancer genomes [101, 102], and continued work is needed with the increasing availability of cancer whole genome data. Encouraged by these results, we are currently developing a newer version of RegulomeDB, which will provide all the features we used in this challenge, including the allelic scores such as information content change of TF motifs. We will also make our prediction scores available to general users, thus to help research on prioritizing non-coding variants in various contexts.

2.6 Publication

The study in this chapter has been published in Human Mutation journal [103]: Dong, S., & Boyle, A. P. (2019). Predicting functional variants in enhancer and promoter elements using RegulomeDB.

CHAPTER III

Prioritization of Regulatory Variants with Tissue-Specific Function in the Non-coding Regions of Human Genome

3.1 Abstract

Understanding the functional consequences of genetic variation in the non-coding regions of the human genome remains a challenge. We introduce here a computational tool, TURF, to prioritize regulatory variants with tissue-specific function by leveraging evidence from functional genomics experiments, including over three thousand functional genomics datasets from the ENCODE project provided in the RegulomeDB database. TURF is able to generate prediction scores at both organism and tissue/organ-specific levels for any non-coding variant on the genome. We present that TURF has an overall top performance in prediction by using validated variants from MPRA experiments. We also demonstrate how TURF can pick out the regulatory variants with tissue-specific function over a candidate list from associate studies. Furthermore, we found that various GWAS traits showed the enrichment of regulatory variants predicted by TURF scores in the trait-relevant organs, which indicates that these variants can be a valuable source for future studies.

3.2 Introduction

Characterizing the biological impact of variation in the non-coding regions of the human genome remains a challenge in the interpretation of human diversity. Genome-wide association studies (GWAS) have identified millions of genetic variants that are associated with diverse disease traits [60]. Most of these variants (∼90%) map to the non-coding regions of human genome [61]. Due to the lack of understanding of these regulatory elements within non-coding regions, it is important to assess the functional consequences of these disease-related variants from GWAS.

To facilitate studies of non-coding genomic regions, large consortia, including ENCODE [104, 105] and the Roadmap Epigenomics projects [41] have defined the human regulatory landscape using high-throughput functional genomics assays. For example, DNase-seq locates open chromatin regions in the genome [5], while ChIP-seq identifies chromatin modification patterns and transcription factor (TF) binding sites within regulatory elements [106, 107, 108]. With further incorporation of variant genotypes into these methods, variants associated with differential TF binding and chromatin states have been described [109, 50]. In addition, massively parallel reporter assays (MPRA) identify regulatory variants that affect gene expression levels directly [25, 26, 29]. These studies demonstrate that a significant number of variants drive regulatory state variation across the population, and potentially explain the diversity in disease risk and phenotype observed from GWAS studies.

Computational tools have helped prioritize regulatory variants in non-coding regions by leveraging knowledge from functional genomics assays. Prediction scores of functional probability for variants are available from tools including RegulomeDB [62], GWAS3D [87], HaploReg [63], DeepSEA [72], DeepBind [89], DanQ [90] and

Basenji [92]. The process of narrowing down a candidate list of variants using these prediction scores can reduce time-consuming validation experiments. However, most current computational tools overlook the uniqueness of gene regulatory networks found within different tissues by only providing a prediction score at an organism level. This can be misleading for research groups focused on tissue-specific functional variants. New tools have recently become available that provide tissue-specific prediction scores, such as FUN-LAD [68], GenoNet [70], cepip [57] and GenoSkyline [69]. However, they mainly utilize epigenetic data from the Roadmap Epigenomics project making it hard to leverage their results against other tissues not included in the Roadmap project [41]. The ENCODE project currently houses thousands of ChIP-seq and DNase-seq datasets in over 200 tissues and cell types, including those from Roadmap project, that can further increase the scale and accuracy of tissue-specific function prediction.

Here we introduce a computational tool, TURF (Tissue-specific Unified Regulatory Features), that prioritizes regulatory variants in the non-coding regions of the human genome. TURF is built on our RegulomeDB framework to allow for easy delivery of our predictions as well as constant updates in the functional annotations across the human genome. We extend our previous algorithm SURF [103] to predict tissue-specific functional variants in addition to the tool's original generic context at an organism level. To construct a high-quality training set, we called 7,530 allele-specific TF binding (ASB) single nucleotide variants (SNVs) in 6 cell lines from over 600 ChIP-seq datasets. We then trained a random forest model using features from functional genomic annotations across all available tissues from ENCODE. This classifier greatly improves the robustness of RegulomeDB v1.1 ranking scores and surpasses other top-performing tools on an independent MPRA dataset.

We then incorporated annotations of histone marks and open chromatin regions in a particular tissue to train a separate random forest model and obtain a final tissue-specific score. The tissue-specific score leverages information from other tissues, as well as retaining the uniqueness of individual tissues. Moreover, we extended the tissue-specific scores to organ-specific scores in the 51 organs with available genomics data from the ENCODE project. The pre-calculated organ-specific scores for all GWAS SNVs from the GWAS Catalog are available at https://github.com/Boyle-Lab/RegulomeDB-TURF and TURF is currently being integrated into RegulomeDB v2.0.

3.3 Methods

3.3.1 Training dataset generation

We identified 7,530 allele-specific transcription factor (TF) binding (ASB) SNVs in 6 cell lines (GM12878, HepG2, A549, K562, MCF7 and H1hESC), which are defined as variants that result in stronger binding of a TF to one allele at heterozygous sites in an individual (Table 3.1). The *AlleleDB* protocol was used to call ASB SNVs [52].

Table 3.1: Number of allele-specific TF binding (ASB) training SNVs in 6 cell lines.

Cell type name	# of ASB SNVs	# of control SNVs
GM12878	1,848	12,382
A549	215	1,357
H1hESC	1,464	6,525
HepG2	2,717	24,479
K562	767	12,635
MCF7	653	8,034

The SNVs in GM12878 and H1hESC were obtained from the 1000 Genome Project [110] and NCBI GEO database (accession number: GSE52457) separately. For the other four cell lines, variants were called from their whole genome sequencing data (data accessible at NCBI SRA database with accession numbers: DRX015191, SRX2598759, SRX285595 and SRX1705314) by *HaplotypeCaller* from the Genome Analysis Toolkit (GATK) v3.6 following GATK's Best Practices [111]. Their diploid personal genomes were constructed using *vcf2diploid* v0.2.6 to avoid alignment biases favoring reads containing reference alleles by mapping to maternal and paternal genomes separately [50]. Copy number variation regions with a read depth of < 0.5 or > 1.5 called from *CNVnator* v0.3.3 [112] were filtered out.

The *AlleleDB* pipeline was run on 864 ChIP-seq datasets in the 6 cell lines from the ENCODE project. In addition to the standard steps in *AlleleDB*, our ASB set was refined by performing beta-binomial tests only within reads overlapping their corresponding TF binding peaks called from the same ChIP-seq dataset. In total, 7,530 ASB SNVs were identified from 638 ChIP-seq datasets.

The ASB SNVs were treated as positive examples in our random forest model. To generate a comparable negative set, we included SNVs from three sources: 1. The 55,611 non-allelic TF binding SNVs, defined by having equal ChIP-seq read counts on two alleles at heterozygous site. 2. The closest variants from each of the SNVs in positive set and outside ChIP-seq peaks (6,373 unique variants in total). 3. A randomly selected set of 1000 genome variants scoring no hits on functional annotations from RegulomeDB v1.1. Those three negative sets were combined and weighted equally in our model.

3.3.2 Building random forest models

For TURF generic scores, seven binary and eight numeric features were created for each variant in the training set (Table 3.2). The seven binary features represent if the variant position overlaps corresponding functional genomic regions by querying RegulomeDB 2.0. Custom scripts were written to retrieve annotations from the RegulomeDB web server. The maximum information content change from PWM was calculated based on the query. Quantiles and variations in ChIP-seq signals pre-calculated from all available bigwig files in ENCODE and prediction scores from DeepSEA were also incorporated. A random forest model was trained to make predictions on the probability of a query variant being functional. The *scikit-learn* 0.20.3 python package was used to train the random forest model, setting the number of trees to 500.

For TURF tissue-specific scores, a separate random forest model was built with 7 binary tissue-specific features (see feature list in Table 3.2). When training with each ASB cell line, the ASB SNVs in the corresponding cell line were labeled as positive variants, while the other variants were labeled as controls. The *scikit-learn* 0.20.3 python package was used, setting the class_weight option as 'balanced'.

3.3.3 Generic scores performance assessment

We evaluated our generic model performance on an independent dataset from an MPRA assay in GM12878 [26]. The labels of the MPRA variants (435 positive variants, 2670 control variants) and prediction scores from DeepSEA [72] and regBase were downloaded from regBase database [77]. The performance of different tools was assessed on the Area Under ROC Curve (AUROC) and the Area Under Precision-Recall Curve (AUPR).

Table 3.2: Feature list in random forest models

Generic features	
TF binding sites from ChIP-seq	Binary variable
DNase I hypersensitive sites from DNase-seq	Binary variable
DNase footprints	Binary variable
DNase footprints with matched TF ChIP-seq peaks	Binary variable
TF motifs from PWM matching	Binary variable
TF motifs from PWM matching with matched TF ChIP-seq peaks	Binary variable
Information content change of two alleles in PWM matching	Numerical variable
Information content change of two alleles in PWM matching with matched TF ChIP-seq peaks	Numerical variable
eQTLs	Binary variable
Quantiles (25%,50%,75% and 100%) and variance of ChIP-seq signals across all available ChIP-seq experiments from ENCODE	Numerical variables
Functional score from *DeepSEA*	Numerical variable
Tissue-specific features (in final ensemble model)	
H3K4me1 peaks from ChIP-seq	Binary variable
H3K4me3 peaks from ChIP-seq	Binary variable
H3K27ac peaks from ChIP-seq	Binary variable
H3K36me3 peaks from ChIP-seq	Binary variable
H3K27me3 peaks from ChIP-seq	Binary variable
DNase I hypersensitive sites from DNase-seq	Binary variable
DNase footprints	Binary variable

3.3.4 Tissue-specific scores performance assessment

The tissue-specific model's performance was evaluated first on three MPRA datasets in GM12878 (E116), HepG2 (E118) and K562 (E123). The labels for the MPRA vari-

ants were obtained from GenoNet (He et al. 2018). The authors labeled the MPRA variants in GM12878 from Tewhey et al. 2018 [26] with a slightly different criteria than regBase [77], resulting in 293 positive variants and 2772 control variants. The MPRA variants in HepG2 and K562 were from [25], where 524 positive variants and 1439 control variants were in HepG2, and 339 positive variants and 1361 control variants were in K562. The same evaluation process as described in GenoNet [70] was used to compare TURF to other available tools, including DeepSEA [72], CADD [88] and GenoSkyline [69]. In detail, we calculated AUROC, AUPR and the correlation coefficient using 1000 replicates of 4:1 random partition of each MPRA dataset. For the divided five parts, four parts were used for training while the remaining part was used for testing.

When evaluating performance with allele-specific TF binding SNVs, pre-calculated scores from GenoNet [70] and GenoSkyline [69] were downloaded.

3.3.5 Extension to organ-specific scores

The mapping from tissues and cell types (i.e. biosamples) to organ names was downloaded from the ENCODE website. When generating organ-specific prediction scores, we combined the annotations from functional genomics data in all biosamples belonging to the corresponding organ. 51/55 organs had available ChIP-seq data of histone marks and DNase-seq data to generate organ-specific scores.

3.3.6 Organ-specific significance scores

We calculated organ-specific significance scores relative to a background set from GWAS variants. The GWAS variants were downloaded and assigned to their mapped traits from the GWAS Catalog [60]. SNVs on chromosomes 1-22 and chromosome X were the only ones considered for the organ-specific scoring. Linkage disequilibrium

(LD) expansion was performed by including SNVs from the 1000 genome project that are in strong LD (R2 threshold of 0.6, precalculated R^2 values downloaded from gs://genomics-public-data/linkage-disequilibrium) with any GWAS SNV. To convert each organ-specific score to a significance score, we calculated the portion of GWAS variants with a greater score in the corresponding organ and did a negative log10 transformation on to the portion.

3.3.7 Organ-specific scores enrichment of GWAS traits

In the enrichment analysis, we focused on the GWAS traits with the enrichment of regulatory variants, which have at least 20 GWAS SNVs and at least 5% of the LD-expanded GWAS SNVs in the trait that have TURF generic scores no less than 0.8 (400 traits in total).

To test the enrichment of organ-specific regulatory variants, each GWAS trait set was first sampled with an equal sized background set from all GWAS SNVs from any trait. Subsequent LD expansion was performed on both the trait set and background set (with a stricter R^2 threshold of 0.8). To reduce the dependencies across SNVs within each set, the SNVs were pruned on each organ individually so that no two SNPs were within 1MB of each other in the same set. Each SNV in decreasing order on organ-specific score was considered, and only retained a SNV if there was no other SNV within 1Mb. After the pruning process, the P value was computed from the Mann–Whitney U test for each organ-trait combination, with the alternative hypothesis as SNVs in the trait set have greater organ-specific scores than the background set. This test was repeated by sampling 100 versions of the background set and a total of 100 P values were obtained for each organ-trait pair. 159 traits had at least one organ passing multiple test correction with an FDR of 5%, applied with the Holm-Sidak test from the python package *statsmodels* v0.12.1.

To define top organ for each trait, overall high scores of the trait were compared to other organs. The negative log-transformed P values from U test were used to compute the z-score of each organ over all 51 organs. The mean z-scores over 100 iterations for each organ-trait pair were calculated and hierarchical clustering on the 51 organs was performed using the ward linkage method. The final heatmap only shows organ-trait pairs with a z-scores mean higher than 0 and passing multiple test correction (FDR threshold of 5%).

3.4 Results

3.4.1 Overview of the TURF algorithm

TURF prioritizes non-coding variants with both generic scores and tissue-specific scores (Figure 3.1). It first uses a random forest model built by training on features from functional genomics annotations in all available tissues and cell types from the ENCODE project [104]. It uses a similar feature set to our previously successful algorithm SURF (Dong and Boyle 2019), including binary features retrieved from the original RegulomeDB ranking scheme and prediction scores from DeepSEA [72]. Furthermore, it includes continuous signals from ChIP-seq assays to increase the resolution of the algorithm (see features list in Table 3.2). Generic scores from the first random forest model predict whether the query variant is functional in any human tissue. Tissue-specificity is further predicted by using a separate random forest model trained on functional genomic annotation features only from a particular tissue. To avoid data availability bias for different tissues, TURF takes advantage of DNase-seq and well-studied histone mark ChIP-seq data that cover most tissues. By combining the probability score from the second random forest model with the generic score from the first model, the resulting tissue-specific score predicts the probability of the query variant being functional in a specific tissue.

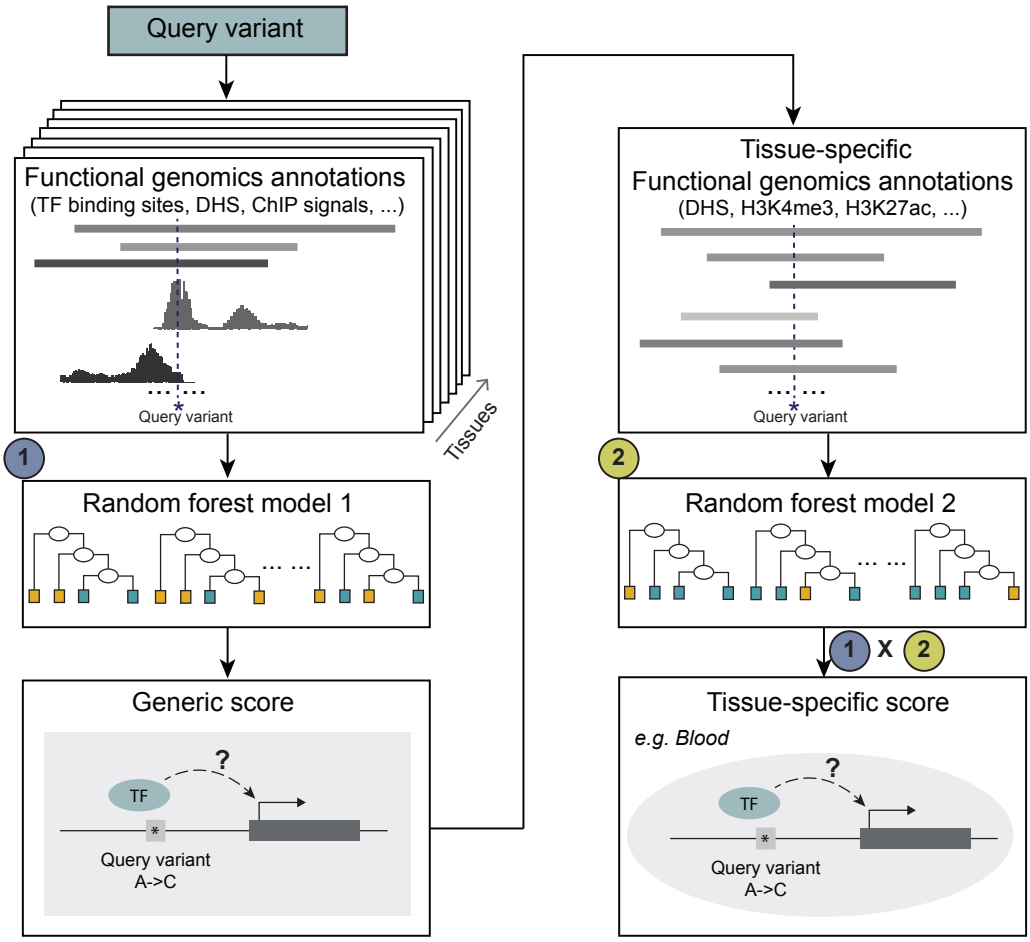


Figure 3.1: Overview of TURF algorithm. TURF generic score predicts the probability of a query variant being functional in any tissue from the first random forest, which used features of functional genomics annotations from all available tissues. By further incorporating annotations from a given tissue, a tissue-specific prediction score is computed by multiplying generic score with the prediction score from a second random forest model

3.4.2 TURF generic score improves the performance of RegulomeDB v1.1 ranking score

TURF improves on the original heuristic ranking score in RegulomeDB v1.1 by providing a probabilistic score generated from a random forest model. By replacing the single empirical decision with sets of decision trees, the model avoids issues caused by excessive reliance on only a few functional genomic annotations. To develop a

training set for the model, we generated a set of variants with high confidence functional confidence through identification of 7,530 allele-specific transcription factor binding (ASB) single nucleotide variants (SNVs) in six cell lines (GM12878, HepG2, A549, K562, MCF7 and H1hESC) by reprocessing 864 ChIP-seq datasets from the ENCODE project using *AlleleDB* v2.0 [52]. ASB SNVs were called if different TF binding affinity with a single nucleotide change at heterozygous sites was observed. We defined a background set using non-allele-specific TF binding SNVs as well as a set of variants outside TF binding regions (see methods).

We evaluated the TURF generic score performance on an independent and orthogonal dataset from a massively parallel reporter assay (MPRA) [26]. This dataset was also utilized as a test set in a previous paper [77], where the authors found DeepSEA scores provided the best prediction model for calling variants functional in tissues. TURF outperformed DeepSEA scores on this MPRA test set with a larger AUROC and the same AUPR (Figure 3.2).

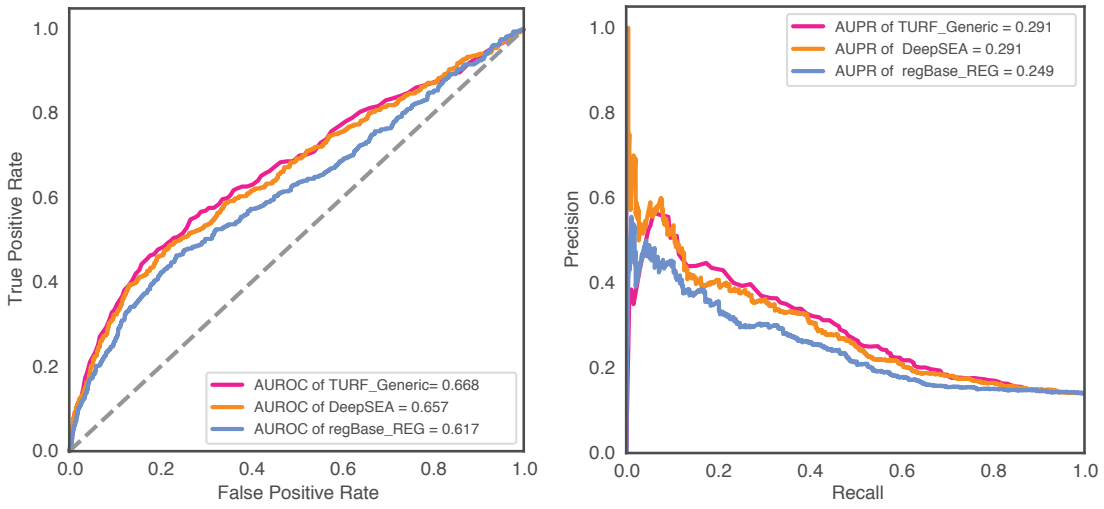


Figure 3.2: TURF generic scores performance on test data from massively parallel reporter assay (MPRA) in GM12878. Performance was evaluated by Area Under ROC Curve (AUROC) and Area Under Precision-Recall Curve (AUPR). 435 positive variants vs 2670 control variants were called in this MPRA validated dataset.

To compare with the original ranking score from RegulomeDB v1.1, we calculated TURF generic scores for all common SNPs from dbSNP153 [113]. The SNPs that originally scored in the highest category, which was largely dominated by eQTL evidence, now show a wider range of scores that better predicted their functionalities, while the overall trend was unchanged (Figure 3.3).

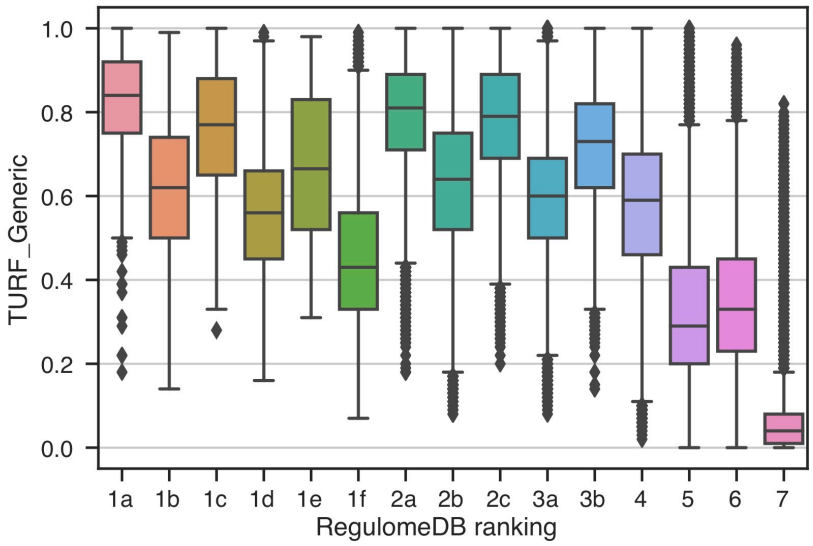


Figure 3.3: Boxplot of TURF generic scores VS RegulomeDB ranking scores on 10,422,004 common SNVs from dbSNP153. X axis represents the original ranking scores from RegulomeDB v1.1, y axis represents the TURF generic scores from random forest model.

3.4.3 TURF tissue-specific scores performance on MPRA data in three cell lines

We further evaluated TURF tissue-specific predictions with MPRA datasets from three cell lines (GM12878, HepG2 and K562) using the same strategy as He et al. [70]. Tissue-specific predictions by TURF had the best performance in GM12878 versus other top performing computational tools (Figure 3.4A and Table 3.3). TURF also has the top AUROC in HepG2 with the second largest AUPR (0.571 compared to 0.572 from GenoNet) and the largest AUPR in K562. Noticeably, the tissue-specific features in the second random forest model have significantly improved the

performance of the TURF generic scores. Among all tissue-specific features, open chromatin regions from DNase-seq in the corresponding cell lines are the most important predictors in all three MPRA datasets. Tissue-specific DNase footprints and active histone marks, including H3K4me2, H3K4me3 and H3K27ac, also play essential roles in variant prediction (Figure 3.5).

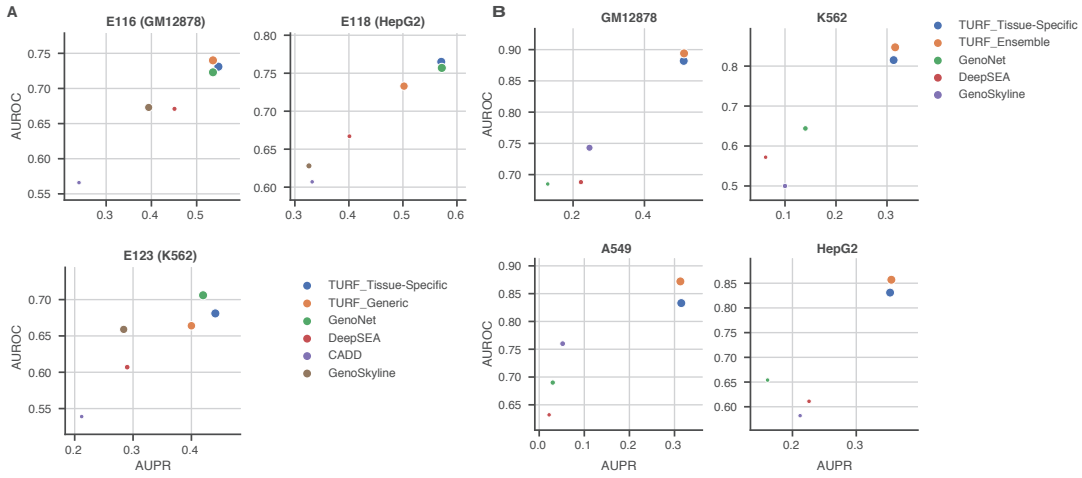


Figure 3.4: Tissue-specific predictions performance comparisons. Each plot shows the AUPR (area under the precision recall curve) on x axis and the AUROC (area under the receiver operating characteristics curve) on y axis. The size of each point represents the pearson correlation. (A) Performance on MPRA data in three cell lines (GM12878: 693 positive variants, 2772 control variants; HepG2: 524 positive variants, 1439 control variants; K562: 339 positive variants, 1361 control variants). (B) Performance on allele-specific transcription factor binding SNVs (see the number of variants in Table 3.1.

Table 3.3: Comparison of performance on tissue-specific predictions for MPRA variants.

	GM12878 (E116)			**HepG2 (E118)**			**K562 (E123)**		
	AUPR	AUROC	COR	AUPR	AUROC	COR	AUPR	AUROC	COR
TURF_Tissue-Specific	0.548	0.731	0.45	0.571	0.765	0.423	0.441	0.681	0.344
TURF_Generic	0.536	0.74	0.447	0.502	0.733	0.372	0.4	0.664	0.294
GenoNet	0.536	0.723	0.442	0.572	0.757	0.429	0.42	0.706	0.34
DeepSEA	0.451	0.671	0.115	0.401	0.667	0.166	0.29	0.607	0.099
CADD	0.24	0.566	0.082	0.332	0.607	0.152	0.212	0.539	0.046
GenoSkyline	0.394	0.673	0.352	0.326	0.628	0.225	0.284	0.659	0.274
TURF_ensemble (Trained from ASB SNVs)	0.484	0.724	0.393	0.479	0.728	0.357	0.329	0.676	0.253

Figure 3.5: Pearson correlation of labels and tissue-specific features in three MPRA datasets (E116: GM12878; E118: HepG2; E123: K562). Blue bars represent positive correlations while red bars represent negative correlations.

3.4.4 TURF tissue-specific predictions on allele-specific TF binding (ASB) SNVs

Despite the power of using MPRA datasets as training sets, they are currently limited in terms of the number of tested variants and the variety of tissues. To obtain a more robust tissue-specific model, we called allele-specific TF binding (ASB) SNVs from 6 cell lines. When trained on ASB SNVS, our tissue-specific models greatly outperformed other methods (Figure 3.4B and Table 3.4). Among the tissue-specific features, DNase-seq peaks and several active histone marks, such as H3K4me2 and H3K27ac, were important predictors of tissue-specific functional variants, similar to what was observed in the MPRA datasets (Figure 3.6). However, DNase footprints

show more variation in feature importance ranking within the 6 cell lines. This indicates the diversity of DNase-seq data quality in different cell lines, and suggests that utilization of a more robust model to compensate for this variation is needed when extending to other tissues not used in the training data.

Table 3.4: Comparison of performance on tissue-specific predictions for ASB SNVs.

	GM12878			K562			A549			HepG2		
	AUPR	AUROC	COR	AUPR	AUROC	COR	AUPR	AUROC	COR	AUPR	AUROC	COR
TURF_Tissue-Specific	0.51	0.882	0.537	0.313	0.815	0.366	0.315	0.833	0.369	0.353	0.831	0.401
TURF_Generic	0.284	0.822	0.39	0.193	0.782	0.301	0.158	0.796	0.256	0.221	0.797	0.323
GenoNet	0.129	0.685	0.158	0.14	0.644	0.113	0.03	0.69	0.094	0.161	0.654	0.134
DeepSEA	0.222	0.688	0.19	0.062	0.572	0.054	0.022	0.632	0.049	0.226	0.611	0.138
GenoSkyline	0.246	0.743	0.34	0.09	0.628	0.144	0.052	0.76	0.119	0.212	0.582	0.133
TURF_ensemble	0.511	0.894	0.538	0.316	0.847	0.376	0.313	0.872	0.369	0.355	0.857	0.406

We then trained an ensemble tissue-specific model using the average predictions from 6 models with feature weights individually learnt from 6 ASB cell lines. The histone mark features were restricted to 5 histone marks that ranked high in feature importance, and had available datasets covering most tissues (i.e. H3K27ac, H3K36me3, H3K4me1, H3K4me3 and H3K27me3). The ensemble model outperformed the individual tissue-specific models when predicting ASB SNVs (Figure 3.4B and Table 3.4). Moreover, this ensemble model trained on ASB SNVs performed better than most of the other tools when tested on the previous independent MPRA datasets in all three cell lines. The exception was GenoNet, which used labels from the MPRA datasets in their training step (Table 3.4). Predictions were computed from this ensemble tissue-specific model on the ASB SNVs in 6 cell types and most exhibited the highest prediction scores in their corresponding functional cell line (Figure 3.7). However, HepG2 ASB SNVs had the least enrichment of high HepG2-specific scores, perhaps due to DNase-seq noise in the dataset as only 25% were in

Figure 3.6: Pearson correlation of labels and tissue-specific features in 6 ASB datasets. Blue bars represent positive correlations while red bars represent negative correlations.

DNase peaks. Some H1hESC ASB SNVs had high scores in K562 and MCF7, implying that a many stem cell regulatory variants are involved in regulation of pathways in differentiated cell lines.

3.4.5 Extension of TURF tissue-specific scores to organ-specific scores

To expand the scale of prediction for TURF, we leveraged tissue-specific functional genomic annotations of tissues belonging to the same organ and generated

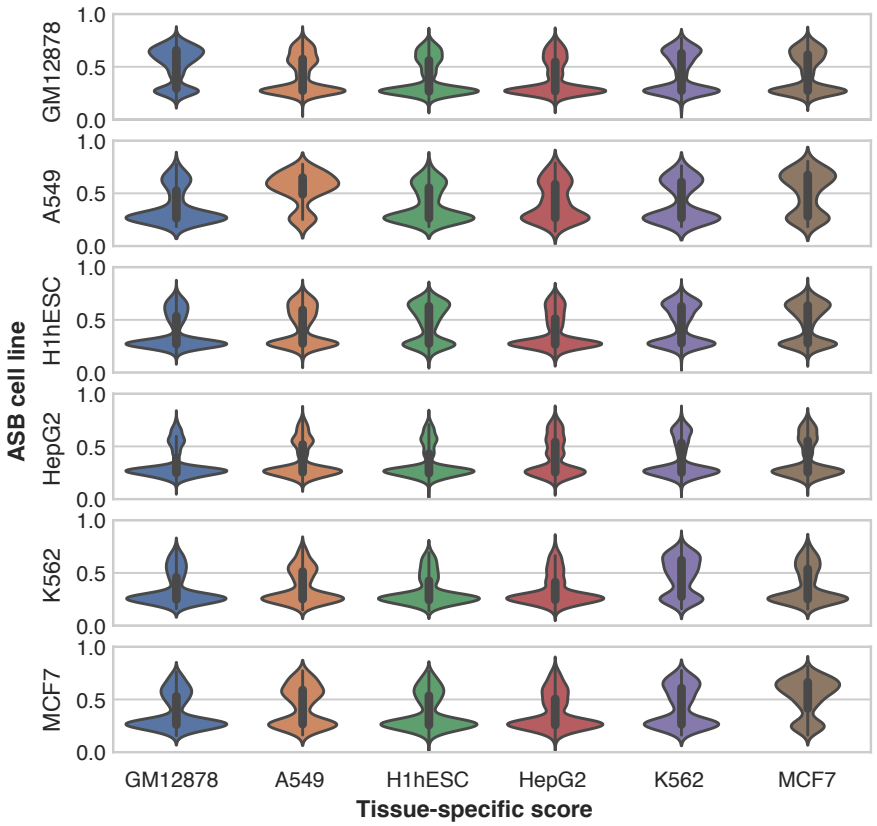


Figure 3.7: TURF tissue-specific scores on allele-specific transcription factor binding (ASB) SNVs called from 6 cell lines. The ASB cell line represents the functional tissue for ASB SNVs in each row. The tissue-specific scores are shown in violin plots with a given cell line in each column. ASB SNVs have overall the highest tissue-specific scores in their functional cell line.

combined organ-specific scores across 51 organs. We were able to recover the organ-specific function of some well-studied regulatory variants in specific genomic loci with TURF scores. For example, TURF's organ-specific scoring was able to pick out the regulatory SNP rs12740374 that affects liver-specific *SORT1* gene expression levels in the 1p13 cholesterol locus [2] (Figure 3.8). The liver-specific function of rs12740374 was also validated in HepG2 MPRA assays [27]. The position of rs12740374 overlaps several active histone mark peaks from ChIP-seq (H3K27ac, H3K4me3 and H3K4me1) and DNase peaks in liver tissues. These multiple lines of genomics evidence prioritized rs12740374 as the top SNP for liver-specific scores within a list of

candidates from previous association studies. In addition to liver, rs12740374 has a high significance score in other organs relevant with cholesterol metabolism, such as adipose tissue and gonad. As another example, TURF also detected a regulatory SNP at the *GATA4* locus in the heart (Figure 3.9) that was initially discovered in a genome-wide association scan on 466 bicuspid aortic valve cases [114].

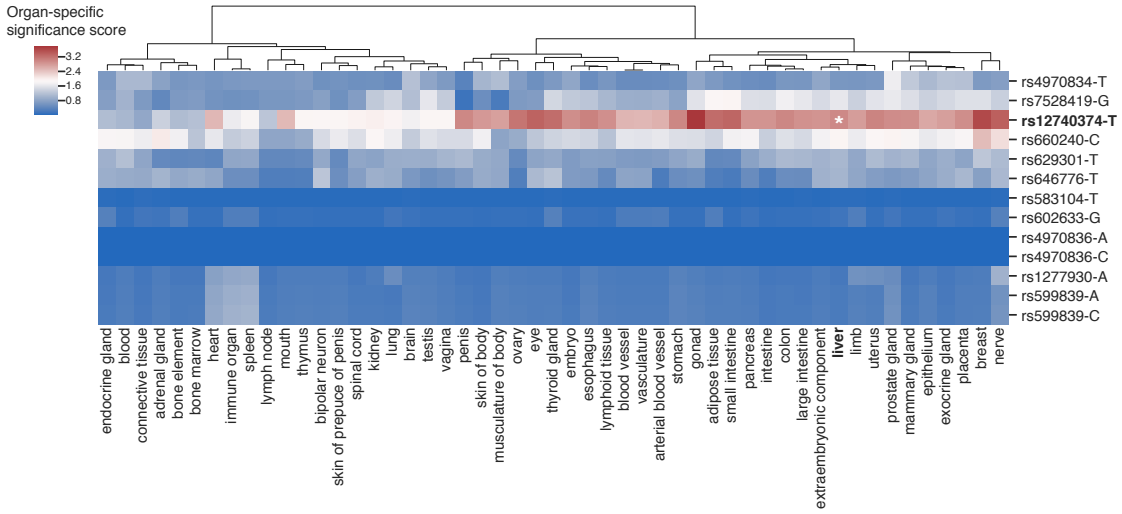


Figure 3.8: Organ-specific significance scores of variants in the 1p13 cholesterol locus. rs12740374 has the top liver-specific significance score compared to other nearby candidate SNPs from association studies, which was validated to affect gene expression level in liver tissue. The organ-specific significance scores were calculated relative to a background set from GWAS variants (See methods).

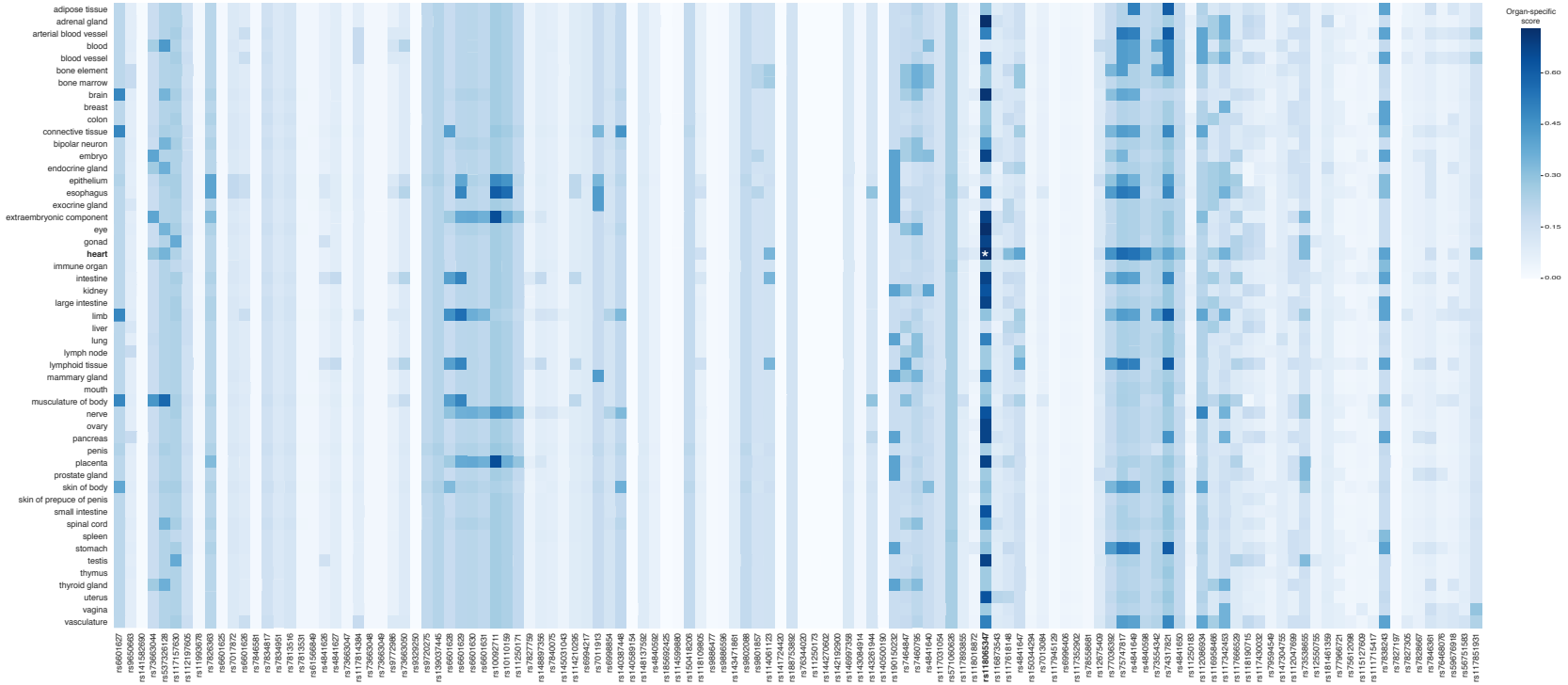


Figure 3.9: Organ-specific scores of variants in the *GATA4* locus. rs118065347 has the top organ-specific score in heart compared to other candidate SNPs found through genome-wide association scan on bicuspid aortic valve cases. rs118065347 was shown to have hear-specific regulatory functionality from a previous study [114].

3.4.6 TURF organ-specific scores prioritize genetic variants associated with traits in relevant organs

We examined TURF organ-specific scores on variants identified from genome-wide association studies (GWAS) using the GWAS Catalog portal [60]. GWAS variants were found to be enriched in regulatory elements of non-coding regions [115, 22]. We tested the enrichment of putative regulatory variants prioritized by TURF scores for a variety of traits. For each trait, the top organ with the highest z-score showed the most significant enrichment of organ-specific regulatory variants relative to the background set from all traits within the GWAS catalog, as well as 50 other organs with the same trait (Figure 3.10 and see full plot in Figure 3.11).

The top enriched organs from diverse traits were consistent with current trait-relevant organ knowledge. For example, many immune system related diseases, such as autoimmune disease, celiac disease and chronic lymphocytic leukemia, showed a high enrichment for regulatory variants functional in immune-related organs, including immune organ, spleen, and lymph node. Traits of immune cells, such as leukocyte, eosinophil and platelet, were also enriched in immune organs. Cardiac traits, including PR interval, which is a measurement in electrocardiography, and coronary artery disease, were enriched in heart and arterial blood vessel. Enrichment in the colon and immune-related organs was demonstrated for Crohn's disease and ulcerative colitis, both inflammatory bowel diseases. Furthermore, several traits of measurement were enriched for organs involved in relevant metabolic pathways, such as cholesterol measurement in liver, apolipoprotein A1 measurement in small intestine [116], renin-angiotensin system (RAS) use measurement in adrenal gland, and alcohol consumption measurement in exocrine gland (i.e. salivary gland). Of note, the enrichment of variants in some traits could be affected by cofactors, such as

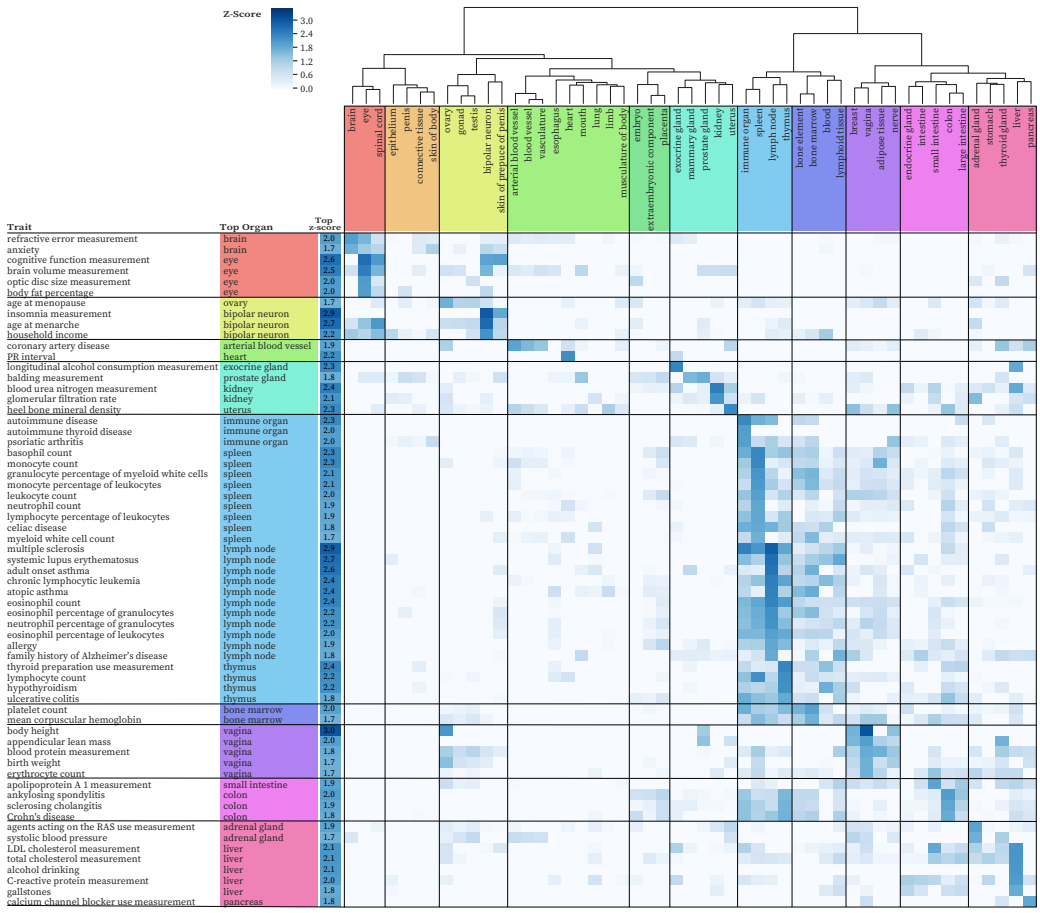


Figure 3.10: Enrichment of regulatory variants with high organ-specific scores over variants associated with diverse traits (z-scores cutoff at 1.7). The z-scores of organs (column) for a given trait (row) are shown. The organ with the highest z-score for each trait is shown in additional columns on left. Only organ-trait pairs with z-scores higher than 0 and passing multiple test correction (FDR threshold of 5%) are shown. Traits with top z-scores < 1.7 were ignored in this plot. See full plot in Figure 3.11.

gender for body height enrichment within the vagina and ovary. Also, some organs seem to share similarities in gene regulatory networks, partly due to overlapping of tissues, or tissues with similar functions across different organs. This explains a mixture of brain and optic traits enriched in either brain or eye, as the optic nerve gene expression pattern was found to be similar to brain tissue [117].

The most enriched organ for potential regulatory variants provides new directions

for understudied diseases or traits. For instance, drugs of calcium channel blockers were found to increase the risk of pancreatic cancer in post-menopausal women [118], while the underlying mechanisms remain unclear. Interestingly, pancreas was the top organ for the calcium channel blocker use measurement trait, which indicates an enrichment of putative regulatory variants functional in pancreas. Thus, additional studies on top variants prioritized by TURF pancreas-specific scores may help further explain the association between pancreatic cancer risk and the use of calcium channel blocker drugs. Similar workflow can be applied to other diseases, such as Alzheimer's disease in immune organs, to determine the causal variants in non-coding regions.

3.5 Discussion

In this study, we developed TURF, a computational tool that prioritizes variants in non-coding regions. Evidence was incorporated from various functional genomic assays to produce robust predictions that were verified via MPRA assays in both generic and tissue-specific contexts. The workflow was designed to identify regulatory variants from association studies with tissue/organ-specific regulatory function. Moreover, we found GWAS variants were enriched with regulatory variants predicted by TURF organ-specific scores in trait-related organs.

To balance between prediction accuracy and data availability, we trained TURF on ASB SNVs identified from ChIP-seq to determine the weight of features in a tissue-specific context, then extended the scale of annotation to an organ-specific level. The TURF tissue-specific scores leverage information gained from other tissues while retaining the uniqueness of the gene regulatory network in individual tissues. We were able to prioritize putative organ-specific regulatory variants across 51 organs in diverse pathways. A number of computational tools have been de-

veloped recently for similar purposes however, most focus on genomic assays and tissues from the Roadmap project [70, 69]. This makes it difficult to utilize their results for tissues not included in the Roadmap project. As an alternative, we took advantage of over 3,000 genomic assays in more than 200 tissues and cell types available from the ENCODE project, expanding the annotation scope and enhancing the robustness of our predictions. Most relevant organs of various GWAS traits were recovered from the organ-specific scores, including some well-studied traits, such as LDL cholesterol measurement and immune diseases. These results were mirrored in active histone marks using epigenomics data from the Roadmap project [22]. In addition, we observed novel organ-trait pairs, including pancreas in calcium channel blocker use measurement, which can help elucidate underlying disease mechanisms. As more functional genomics datasets are generated, our algorithm is flexible allowing for addition of new tissues by querying histone mark and DNase features within the new tissue and then computing new tissue/organ-specific scores.

Despite the large scale of annotation utilizing the 51 ENCODE organs, further refinement of the organ terms and the tissues assigned to each organ is possible. Some traits in 3.8 showed enrichment in non-relevant organs, such as household income in the bipolar neuron. This could be partly due to cofactors within individual GWAS samples, but can also imply an imbalance in the number of genomic datasets across diverse organs as the bipolar neuron (i.e. ear) only contains one ENCODE biosample. Due to the limitation of data availability, we only used 7 tissue-specific binary features when building the second random forest model. With more functional genomics data being generated, especially those targeting more histone marks, we can expand our feature set and generate a wider spectrum of prediction scores. The organ-specific scores can then be normalized across different organs to eliminate bias from data

availability. The organ-specific scores for a variant will be more comparable over a list of interested organs.

We used MPRA data to validate our method as these assays provide more direct evidence of variants affecting gene expression than other association analyses, such as eQTLs, which can be affected by variants that are in strong linkage equilibrium. However, we could only test our model in three MPRA cell lines when comparing performance to other tools. We found one tool used MPRA data labels causing overfitting when tested on ASB SNVs. We built an ensemble model trained on SNVs from 6 cell lines to avoid the overfitting. With more MPRA data becoming available in the future, we can provide a more thorough comparison of performance and further refine our model by including training variants from more cell types or more types of assays.

Overall, TURF is able to prioritize regulatory variants with either generic or tissue-specific functions. We expect our tool to enhance future studies on functional consequences of regulatory variants associated with diseases from GWAS. The organ-specific scores generated here will be incorporated into the RegulomeDB database soon, making it a useful tool for broad communities.

3.6 Publication

The manuscript of the work in this chapter has been submitted and is accessible in bioRxiv [119]:

Dong, S., & Boyle, A. P. Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome.

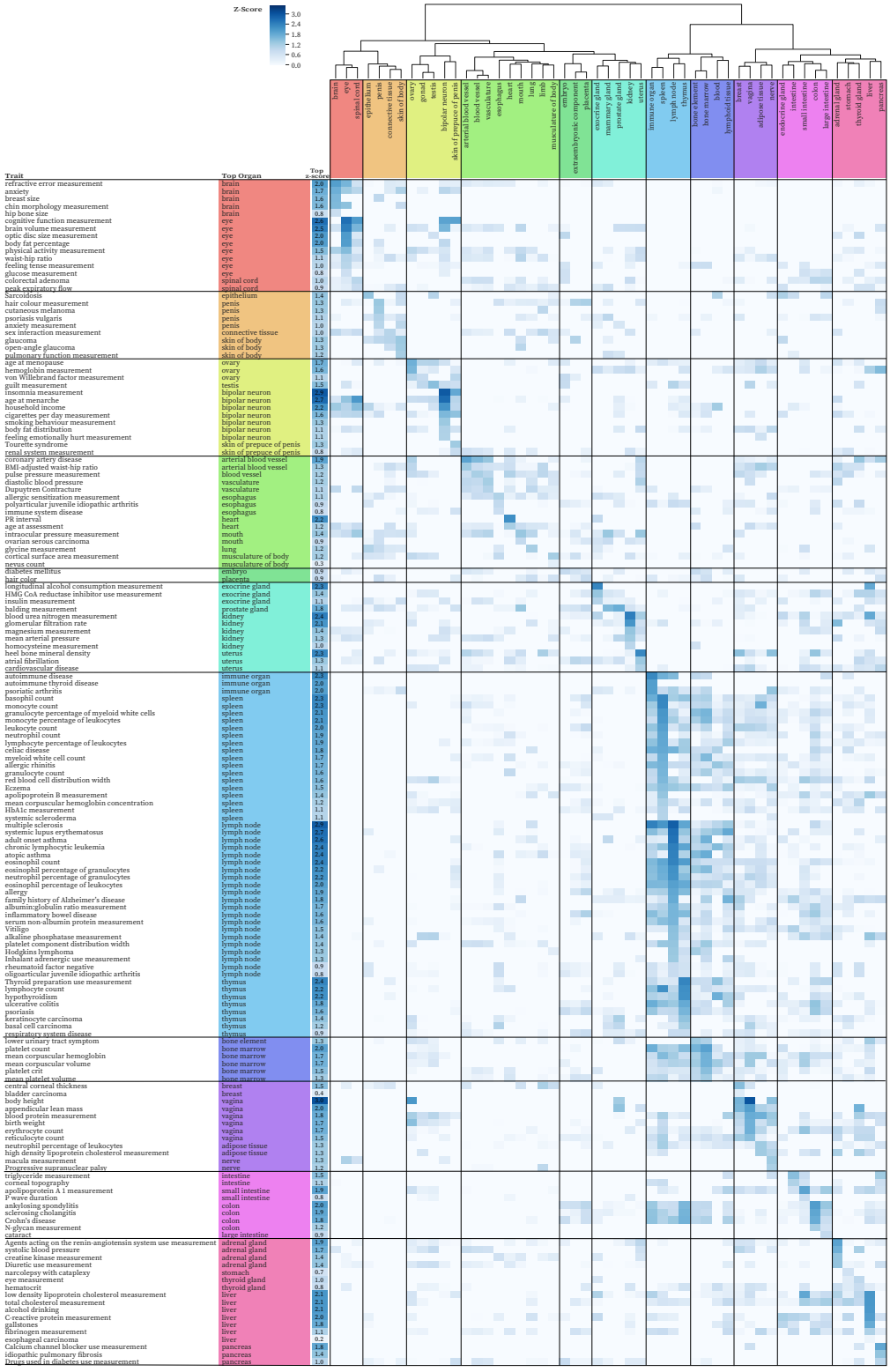


Figure 3.11: Enrichment of regulatory variants with high organ-specific scores over variants associated with diverse traits (full plot). The z-scores of organs (column) for a given trait (row) are shown. The organ with the highest z-score for each trait is shown in additional columns on left. Only organ-trait pairs with z-scores higher than 0 and passing multiple test correction (FDR threshold of 5%) are shown.

CHAPTER IV

RegulomeDB 2.0: An Online Tool for Non-Coding Variant Annotation

4.1 Abstract

Nearly 90% of the disease-associated variants called from GWAS are located in non-coding regions of the human genome. The functional consequences of those variants remain underexplored in many cases. We developed the RegulomeDB database to address this challenge by providing functional annotations and ranking scores on query variants from users' input. For each query variant, RegulomeDB intersects its position with the evidence from functional genomics experiments, such as ChIP-seq and DNase-seq assays. In this work, we released a new version of the RegulomeDB web server. We have incorporated interactive charts and genome browser views to present more intuitive functional annotations to users. In addition, we extended the annotations by including new datasets, such as those from ENCODE phase 3 and eQTLs from the GTEx project. We also integrated a new probabilistic scoring scheme to provide more robust and accurate predictions than the previous ranking scores. We demonstrate these updates as RegulomeDB 2.0 and illustrate the annotations from the web server's new interface (http://regulomedb.org).

4.2 Introduction

Understanding the biological impact of variants in the non-coding regions of the human genome is a main challenge. Nearly 90% of the disease risk-associated single nucleotide polymorphisms (SNPs) identified from genome-wide association studies (GWAS) are in non-coding regions. And similarly, only 25% of Mendelian disease patients have had their mutations in protein coding regions [120]. The abundance of disease-associated variants in non-coding regions makes it desirable to extend these studies to understand functional consequences.

The annotations from functional genomics assays can provide additional information for variants called from GWAS. For example, when studying a specific SNP identified in a GWAS study, although it has not been recorded in the literature, the regulatory nucleotides in linkage disequilibrium annotated from TF ChIP-seq assays may implicate genes and pathways that contribute to the development of diseases. More broadly, the annotation of non-coding variants requires integrating multiple layers of functional information, including putative regulatory elements broadly identified from high-throughput sequencing datasets (e.g. DNase-seq, ChIP-seq, and ATAC-seq).

Despite the benefit of incorporating functional genomics evidence when examining non-coding variants, the lack of available annotation tools limits the use of such data. The majority of resources for clinical purposes has been focused on coding regions as an application to exome sequencing data [121, 122], which consists only <5% of human variation (International HapMap Consortium 2005; 1000 Genomes Project Consortium et al; International HapMap Consortium et al). More recently, whole-genome sequencing has become more common with less cost and large projects such

as NIH's AllofUs [123], the European '1+ Million Genomes' initiative [124], and the Million Vets Program [125]. This shift from exome to genome leaves a gap in our knowledge as there are many fewer tools accessible for researchers and clinicians to annotate non-coding variants.

We have previously built the RegulomeDB v1.1 database for annotating variants in non-coding regions. Intersecting the query variants with evidence from various functional genomics experiments provides users with regulatory information and ranking scores for all input variants to prioritize putative functional variants [62]. In this work, we updated the RegulomeDB web server to display the query results more intuitively with interactive charts and the genome browser views. We also reworked our pipelines to enable efficient search against numerous genomic regions. We expanded the database to including newly generated data from ENCODE phase 3 and the eQTLs from the GTEx project. Also, we provide a new probabilistic score for each query variant in addition to the original ranking score [103].

4.3 Methods

4.3.1 Data collection and processing

Genetic variants The information of genetic variants was retrieved from dbSNP153 [113], including the positions and allele frequencies. The linkage disequilibrium (LD) was calculated on genotype information from 1000 Genomes Project in five super populations (AFR, AMR, EAS, EUR, and SAS) (gs://genomics-public-data/linkage-disequilibrium).

eQTLs The eQTLs from the GTEx project across 49 human tissues were collected. We performed LD expansion by including variants in strong LD with evariants from eQTLs results (R^2 threshold of 0.8). The evariant-egene pairs with the corresponding tissue were parsed to the database.

PWM matching We downloaded the PWMs (position weight matrices) of 722 non-redundant TF motifs from JASPAR 2020 database [20].The kmers matching to TF motifs were called by TFM P-value with a threshold at 4^{-8} for each PWM. Bowtie was used to map the kmers on genome to determine the final PWM matching positions for the TF motifs [126]. The information content from each PWM was also integrated into the database and used as one feature to calculate the probabilistic score.

Epigenomes The epigenomic data from ENOCDE was directly retrieved from the ENOCDE portal, including the newly generated data from phase 3 (Table 4.1).

Data type	**# of datasets**	**# of genomic intervals**
TF ChIP-seq (ENCODE)	3,494 cell types/conditions	36,913,587
DNase-seq	1,821 cell types/conditions	139,371,262
eQTLs (GTEx)	48 tissues	3,124,345 (SNVs)
Predicted TF binding sites (PWM matching)	722 motifs	123,955,090
FAIRE sites	25	4,816,196
DNase I footprints	50 cell types	128,266,803
dsQTLs	6069 SNPs	6,069
VISTA enhancers	1448 enhancers	1358
Validated SNPs affecting binding	855 SNPs	855
Manual annotations	6 genomic regions	282

Table 4.1: Statistics on database content. Number of datasets under each data type includes all experiments across different treatment condition and biosamples. The bolded data types were updated ones in RegulomeDB 2.0, other statistics were obtained from [62].

Precalculated scores We calculated the ranking scores and the new probabilistic scores from SURF for common variants from dbSNPs. We also calculated the functional significance score from DeepSEA [72] for all possible biallelic variants on genome regions within the union set of DNase peaks from various conditions and biosamples, due to the limitation of computational time for the whole genome. We implemented the score into the back-end of database to allow rapid queries in future.

4.3.2 Database and web server design

RegulomeDB annotates a variant by intersecting its position with genomic intervals identified from a massive number of experiments and computational approaches. The database directly integrates the datasets from ENCODE portal through a limited mirror of the ENCODE system (https://github.org/ENCODE-DCC/regulome-encoded). The genomic intervals are stored in an Amazon S3 bucket with BED formatted files associated with JSON objects containing the information of source experiments and computational pipelines. These BED files are then indexed in *Elasticsearch* (https://www.elastic.co/) as in integer range to enable efficient search against a query position. In total, over 830M genomic intervals are indexed in *Elasticsearch*. After each search, the JSON objects associated with the intersected intervals are returned and passed on to generate ranking scores from RegulomeDB 1.1 and new probabilistic scores from SURF [103]. The query results are displayed with a web interface containing drawing charts and interaction figures, which can be customized by users.

4.4 Results

4.4.1 Usage and interface

The RegulomeDB 2.0 web server accepts any query variant on the whole genome, mainly designed for querying single nucleotide variants approximately up to 500 at one input (Figure 4.1). The input query variant can be in three formats: 1) rsID from dbSNP153; 2) chromosome position for a single nucleotide variant; 3) chromosome position for a chromosome region. In the third case, all variants on the chromosome region at >1% allele frequency from dbSNP153 will be queried. The database then intersects the variant(s) position with the genomic intervals of

annotations from functional genomics experiments and returns a sortable summary table of variant scores (Figure 4.2). The new probabilistic score is integrated into the summary table, which further prioritizes the variants with the same ranking scores. In addition, a dbSNP rsID will link to the query variant if it exists.

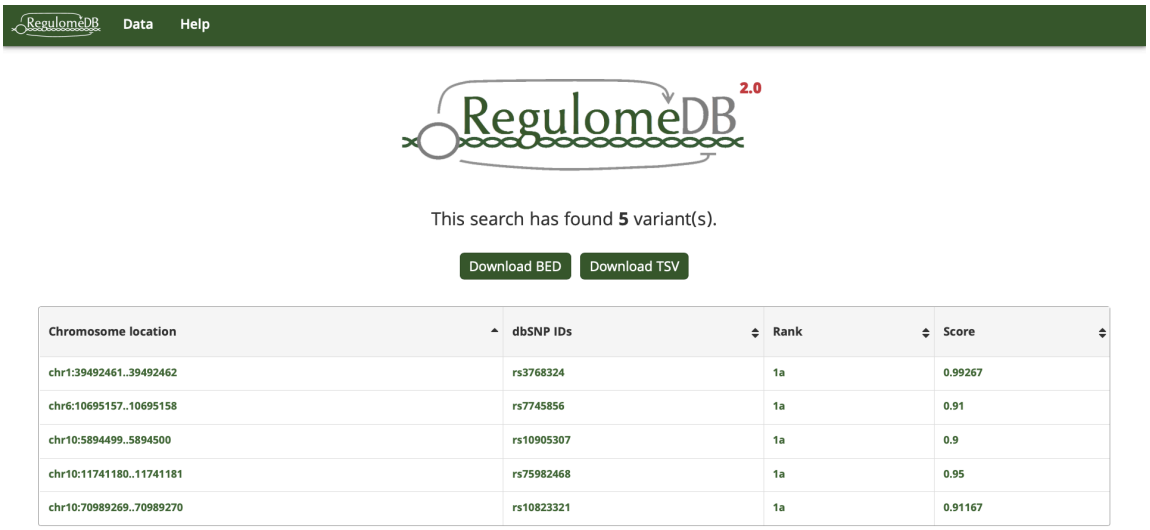
By clicking on any field of a row in the score table, a more detailed information page on genomic evidence is shown for the variant of interest. The top of the page (Figure 4.3) shows some basic information on the variant position, scores, and allele frequencies from dbSNP153. While on the bottom is the initial summary section on genomic annotations' hits (Figure 4.4). Since a single query can hit up to 2,000 results, the initial summary section is divided into 5 data types, including transcription factor binding sites from ChIP-seq, chromatin states from chromHMM, chromatin accessibility from DNase-seq, PWM matching, and QTLs. Furthermore, a genome browser section is also available to assist variant interpretation.

Each of the six sections can be clicked to display more details on the genomic hits from specific assays, such as the biosample of DNase peaks and the transcription factors of ChIP-seq peaks (Figure 4.5). The chromatin state tab shows the chromHMM state for each of the 127 tissues from Roadmap Epigenomics Project (Figure 4.6). Furthermore, the genome browser tab provides an interaction view for exploring the gene transcripts along with DNase-seq and ChIP-seq peaks near the variant of interest (Figure 4.7).

4.5 Discussion

The RegulomeDB 2.0 web server provides a user-friendly tool to annotate and prioritize variants in non-coding regions. The update of our pipeline in processing BED formatted files makes it straightforward to integrate new datasets from func-

Figure 4.1: The RegulomeDB landing page and interface. Users can input rsIDs, variant coordinates, or coordinate ranges. The input of a coordinate ranges will score all common variants within the range.



Chromosome location	dbSNP IDs	Rank	Score
chr1:39492461..39492462	rs3768324	1a	0.99267
chr6:10695157..10695158	rs7745856	1a	0.91
chr10:5894499..5894500	rs10905307	1a	0.9
chr10:11741180..11741181	rs75982468	1a	0.95
chr10:70989269..70989270	rs10823321	1a	0.91167

Figure 4.2: The initial sortable summary table of ranks and new probabilistic scores for all query variants.

tional genomics data. Currently, we have annotations mapped to genome assembly hg19, and we will update them for GRCh38 once all mapped files are ready. We will also integrate new genomic annotations on CRCh38, including the DNase footprints

Figure 4.3: The summary page of a query variant (part 1). The scores and allele frequencies are shown at the top, and a localized region of the genome is shown in the diagram by finding the 10 nearest common SNPs (MAF > 0.05). The query SNP is shown in red.

called from a new pipeline TRACE [9] on the latest ENCODE DNase-seq reads and chromatin states predicted from SEGWAY [24] and ChromHMM [127]. In addition to the current data types, we will incorporate 3D chromatin structures, such as chromatin loops between promoters and enhancers and topologically associated domains (TADs), to provide users with more information on the underlying mechanisms for putative regulatory variants.

We also plan to add new features to the web server, including the tissue-specific function and target gene assignment. In detail, we will calculate the tissue/organ-specific scores from TURF (reference) for each query variant, which needs parsing genomic intervals from histone modification ChIP-seq data in addition to the current set of genomic annotations. We will group the genomic evidence of each query variant based on the underlying organ, which can be retrieved directly from the JSON object

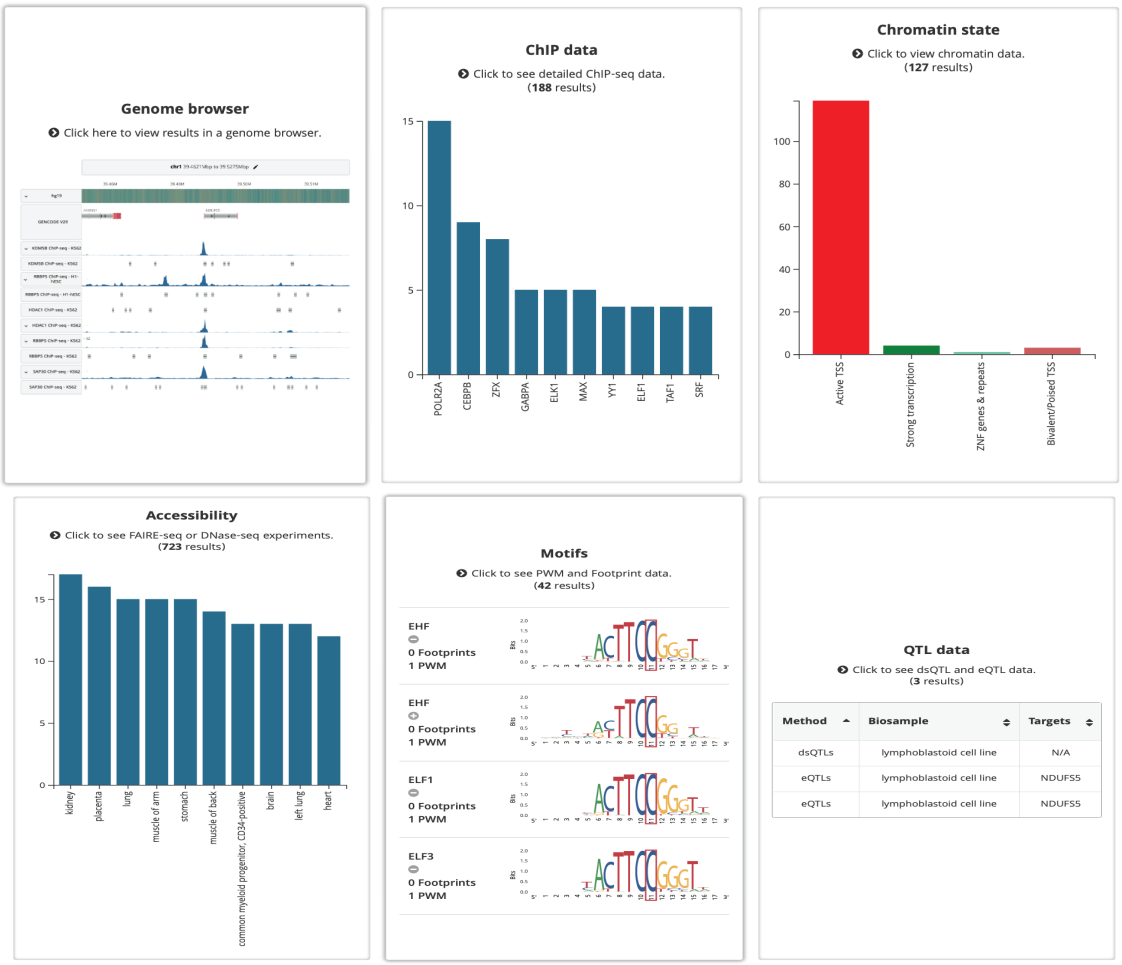


Figure 4.4: The summary page of a query variant (part 2). The genomic evidence is divided into 5 data types, in addition to the genome browser to assist in variant interpretation.

associated with each experiment. On the other hand, we plan to provide functional hypotheses to the putative regulatory variants by assigning target genes. We will use three strategies: 1) the nearest genes to the regulatory variant; 2) the egene mapped to the regulatory variant if it is in the eQTL dataset; 3) the genes within the same TAD of the regulatory variant. These two new features will greatly help researchers study the functional consequences of regulatory variants associated with disease or trait in specific tissues/organs.

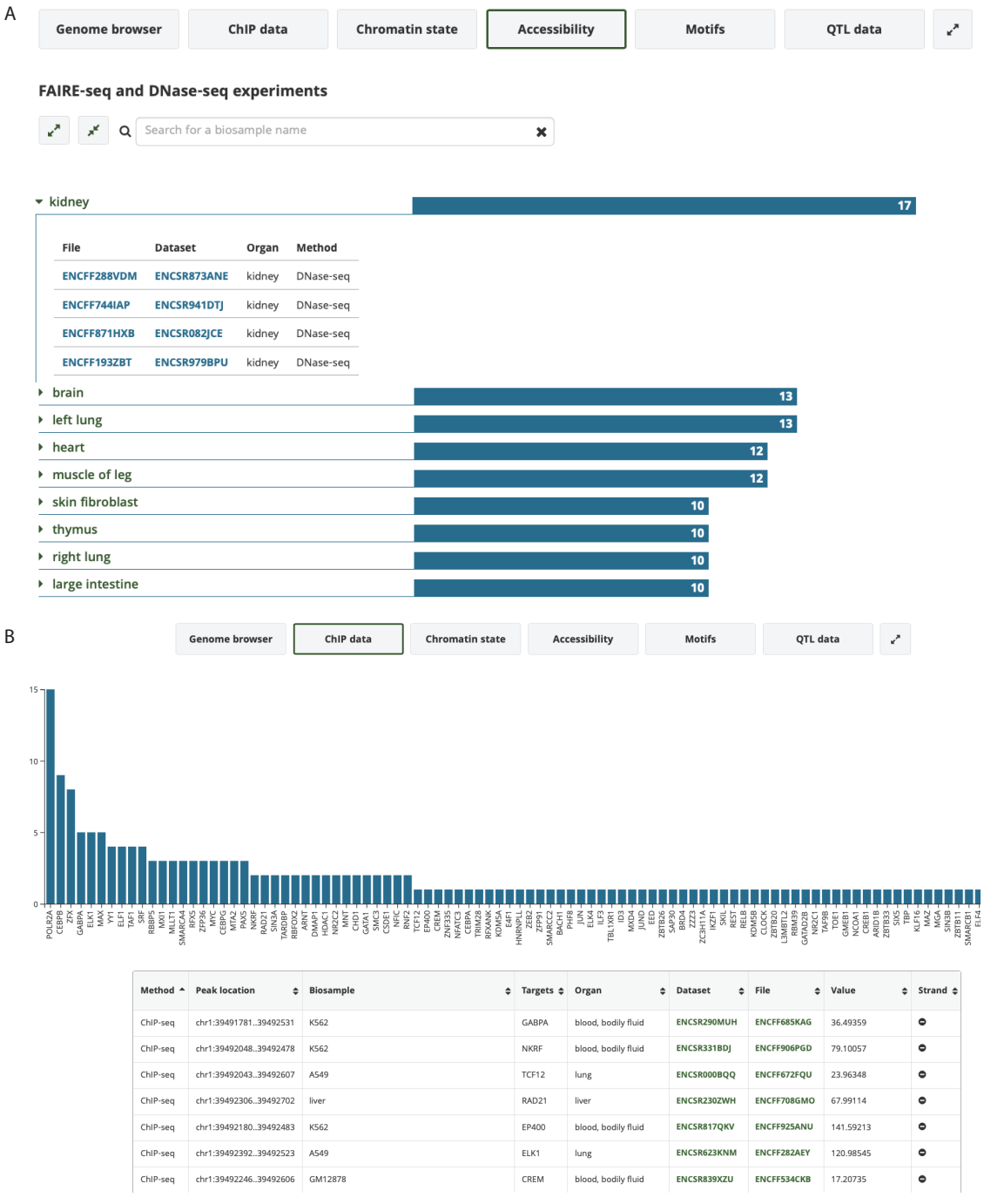


Figure 4.5: The views under chromatin accessibility and ChIP data tabs. The biosamples from DNase-seq assays (A) and the transcription factors from ChIP-seq assays (B) are shown.Each Dataset or File shown in the tables can be clicked to show more information on the corresponding experiment.

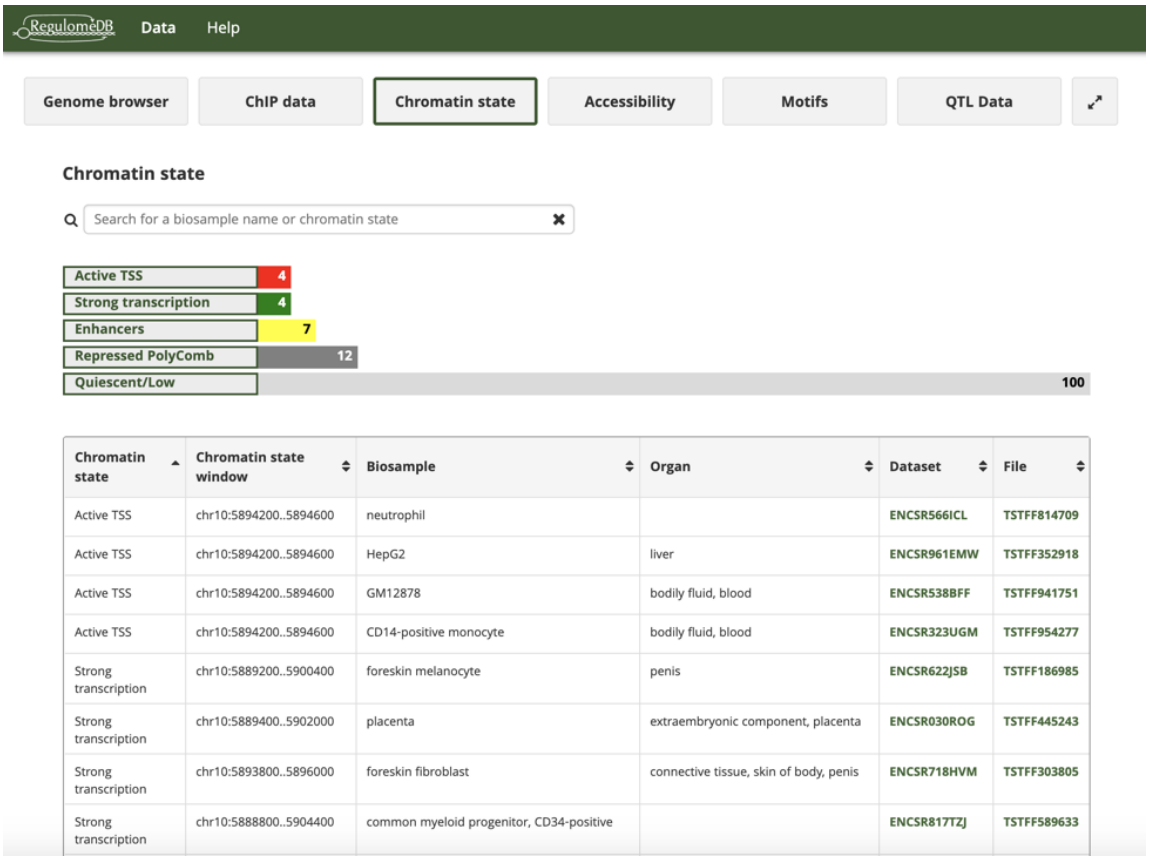


Figure 4.6: The view under chromatin state tab. It shows the chromatin states for each of 127 biosamples from Roadmap Epigenomes Project. Bar graphs display the enrichment of the query variant in each state.

4.6 Publication

The work described in this chapter is being prepared for publication. I will be one of the co-first authors. This work is done in collaboration with Michael Cherry's lab in Stanford University. Yunhai Luo and Benjamin C. Hitz curated data and integrated the ENCODE portal into the database. Emma R. O'Neill designed the interface for the web server. My contribution involved data curation and new probabilistic score integration.

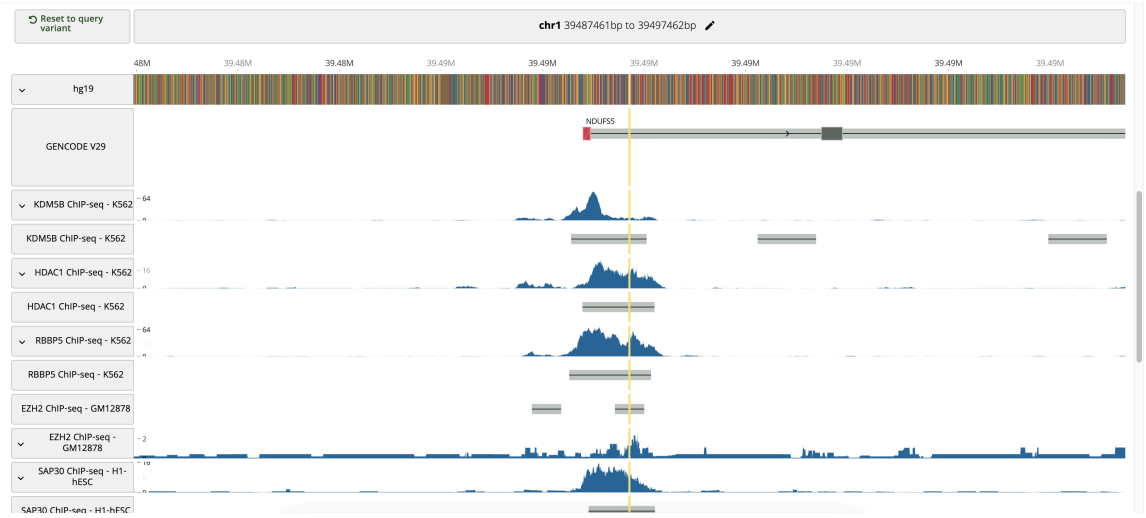


Figure 4.7: The view under genome browser tab. It shows an example variant (the yellow line) near the transcription start site of NDUFS5. Several tracks from ChIP-seq assays are shown here, with both signals from bigwig file and peak regions from bed files. The tracks can be filtered by assay names, biosamples and/or TF targets.

CHAPTER V

Assigning Target Genes for Regulatory Variants with eQTL Studies and 3D Conformation Annotations

5.1 Abstract

Assigning the target genes to putative regulatory variants can provide functional hypotheses and help understand their functions. However, it is challenging because the target genes can be hundreds of kilobases apart from the regulatory elements in the linear distance on the same chromosome. The interactions between regulatory elements and target genes are brought through chromatin looping, which can be mapped by 3D conformation assays such as Hi-C. In addition, association studies such as eQTL analysis also detect such regulation but in an indirect way. In this study, I explored the computational pipelines to assign target genes to ASB SNVs with evidence from Hi-C and eQTL studies in a tissue-specific manner. I present an example of the target gene (UGT2B4) assigned to an ASB SNV (rs7438135) that might help explain the regulatory network involved in lung cancer progression and treatment. The coordination of Hi-C data in various tissues is needed to apply this pipeline in a more general context in the future.

5.2 Introduction

The regulatory variants in non-coding regions can be prioritized by leveraging evidence from functional genomics experiments. Despite recent advances in predicting regulatory variants with computational tools, major challenges remain in interpreting underlying mechanisms for putative regulatory variants. Assigning the target genes to regulatory variants can provide vital information to understand further their functional consequences and association to diseases or traits. One strategy is to assign the gene of which transcription start site is the closest to the regulatory variant in the linear distance on the same chromosome. However, this is an overly simplified model in many cases. The looping between enhancers and promoters can bring the regions up to hundreds of kilobases within spatial proximity. Moreover, the gene regulatory network is dynamic across various cell types and treatment conditions. Therefore, tissue-specific genomic information additional to the linear distance on chromosomes is essential to assign the correct target genes to regulatory variants.

The 3D conformation assays such as Hi-C and Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) have yielded genome-wide chromatin interaction maps [32, 33, 34, 35]. The topologically associating domain (TAD) is a functional structure identified from Hi-C heat maps [80]. These ∼100kb regions are believed to represent gene regulation units since the interaction frequencies between genomic regions within domains are significantly higher than the frequencies outside domains [128]. Therefore, target genes can be assigned to regulatory variants with the assumption that they should locate within the same TAD to accomplish regulatory functions. Meanwhile, chromatin loops are also useful for defining target genes, which represent the long-range chromatin interactions such as promoter-enhancer pairs. The looping

between promoters and enhancers can be identified from the corner dot structures in Hi-C heatmap. Annotations from histone modifications can be incorporated to define the regulatory elements within loop domains. Furthermore, there are assays that more directly capture promoter-enhancer loops, such as ChIA-PET enriched for RNA Polymerase II and promoter-capture Hi-C.

In addition to the assays mapping interaction between genomic regions, association analysis between variant genotypes and quantitative traits from genomic experiments can also help identify target genes to regulatory variants. For example, eQTLs are genomic loci where the genetic variation is statistically associated with variation in gene expression levels, referred to as evariant-egene pairs. Thus, the egenes can be viewed as potential target genes for those evariants. Noticeably, since eQTL studies are confounded by linkage disequilibrium (LD), the leading evariants are often not the causative variants. Therefore, it is necessary to extend the annotation beyond leading evariants by combining LD structures. Numerous datasets on eQTLs and LD structures are available from the Genotype-Tissue Expression (GTEx) project [46, 47]. Up to now, the GTEx project has identified eQTLs in 838 individuals over 49 human tissues, involving over 20,000 genes and 4,000,000 variants. Meanwhile, large consortia, including the 1000 genome project, have identified genome-wide variants with their genotypes and allele frequencies across diverse populations, which are useful in defining LD structures.

As one example of putative regulatory variant, allele-specific transcription factor (TF) binding (ASB) variants show variation in TF binding affinity across two chromosomes on the heterozygous sites within an individual. Thus, they are the candidate regulatory variants affecting TF occupancy, which can be identified through the imbalance on the number of TF ChIP-seq reads mapping to two alleles. Rather

than a simple binomial test on the number of reads from two chromosomes, more factors need to be considered to identify ASB variants, such as the bias from mapping to reference genome, the aneuploidy in cancer cells, and the variation from technical effects. Previous studies have applied statistical approaches such as Beta Binomial models to overcome some of the problems. Thousands of ASB variants in more than 30 cell types have been identified [50, 51, 52, 53, 54, 55]. However, the functions of those variants are underexplored, especially in a tissue-specific context.

In this study, I explored the workflow to assign target genes to regulatory variants by incorporating evidence from TADs and eQTLs. I used the ASB SNVs called from over 600 ChIP-seq datasets in 6 cell lines as an example set of putative regulatory variants. I showed that tissue-specific genomic information can provide functional hypotheses on the ASB SNVs, which can also help refine functional variants relevant to specific pathways.

5.3 Methods

5.3.1 Identification of allele-specific TF binding (ASB) SNVs

The details on calling ASB SNVs were described in Chapter 3 (3.3.1). Briefly, 7,530 ASB SNVs were identified in 6 cell lines from over 600 TF ChIP-seq datasets. The genotypes of variants in each cell line were called from whole-genome sequencing datasets by *HaplotypeCaller* from the Genome Analysis Toolkit (GATK) v3.6 [111]. The copy number variation regions called from *CNVnator* v0.3.3 were removed from this analysis [112]. The AlleleDB pipeline was sued to call ASB SNVs with a Beta Binomial model. The personal genomes on paternal and maternal alleles were built to avoid the bias from mapping to the reference genome. Meanwhile, 55,611 non-ASB SNVs were also called from the definition that there are equal ChIP-seq read counts on the two alleles at heterozygous sites.

5.3.2 Cell type-specific TADs from Hi-C experiments

Four of the ASB cell lines (GM12878, K562, A549, and H1hESC) have available TADs called from Hi-C experiments in their corresponding cell types in hg19 genome assembly (Table 5.1). The TADs regions were downloaded from 3D genome browser [129], which were uniformly called by the same pipeline from Dixon et al. in all cell types [80].

Table 5.1: TADs in four ASB cell lines.

ASB cell line	Reference for TADs	# of TADs	Average bases of TADs
GM12878	GM12878_Rao2014	3,120	842,586
K562	K562_Rao2014	3,019	904,221
A549	A549_ENCODE3-Dekker	1,842	1,411,620
H1hESC	H1-ESC_Dixon2015	2,499	1,065,940

5.3.3 Tissue-specific eQTLs from the GTEx project

eQTLs in 49 human tissues are available through the GTEx project. The most relevant tissue in GTEx was mapped to each ASB cell line, except for the stem cell line H1hESC (Table 5.2). LD expansion was performed to extend the target genes from egenes to the variants that are in strong linkage with evariants. The precalculated R^2 values in the EUR population were downloaded from gs://genomics-public-data/linkage-disequilibrium. On average, each evariant has 1.5 variants in strong LD from 1000 genome project with R^2 threshold of 0.8 (Table 5.3).

5.3.4 Tissue-specific genes

The list of genes with tissue-specific functions was downloaded from a previous paper [130]. The authors compared the gene regulatory networks of TFs and target genes in 38 tissues from GTEx project. The genes in tissue-specific nodes were

Table 5.2: The relevant GTEx tissues for each ASB cell line.

ASB cell line	Cell type name	GTEx tissue
GM12878	Lymphoblastoid	Lymphocytes
K562	Leukemia	Whole blood
MCF7	Mammary cancer cells	Breast mammary
HepG2	Hepatocellular carcinoma	Liver
A549	Lung cancer cells	Lung
H1hESC	Embryonic stem cells	NA

Table 5.3: The number of unique eSNVs from the GTEx project after LD expansion.

GTEx tissue	# of unique eSNVs	# of unique eSVNs after LD expansion (R^2>0.8)
Lymphocytes	181,578	305,794
Whole blood	601,629	897,196
Breast mammary	430,465	648,369
Liver	207,822	325,962
Lung	775,441	1,126,413

identified as genes with tissue-specific functions. Noticeably, the genes can have a multiplicity greater than one, which means they have similar functions in a subset of the 38 tissues, but the function is still distinct from other tissues.

5.4 Results

5.4.1 Comparison of the target genes for ASB SNVs from three approaches

Three approaches were applied to assign target genes for ASB SNVs in 6 cell lines: 1) The gene with the nearest transcription start site (TSS) from the query variant; 2) The gene(s) within the same TAD as the query variant; 3) the egene(s) from eQTL studies where the evariant is in strong LD with the query variant (Figure 5.1). These approaches were all performed in a tissue-specific manner, where only the TADs and eQTLs from corresponding cell types/tissues were used.

The overlapping across target genes from three approaches is shown in Figure 5.2. Overall, the TADs approach assigned the largest number of target genes to ASB

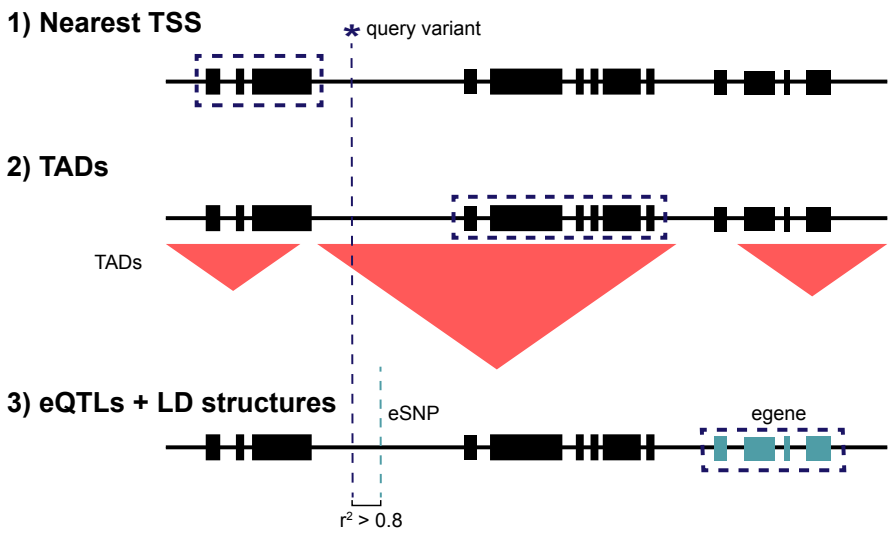


Figure 5.1: Three approaches for target gene assignment. Genes with dashed borders will be assigned as target genes of the query variant with the following three approaches: 1) The gene with the nearest transcript start site (TSS). 2) The gene within the same topologically associating domain (TAD). 3) The egene from eQTL studies of which the eSNP is in strong linkage disequilibrium (LD) with the query variant.

SNVs, with an average of 10 target genes to each ASB SNV. This large number of target genes is primarily due to the low resolution on some of the Hi-C experiments in early studies. While this can indicate high false positive rates, it can still be useful when a low false negative rate is preferred to return all possible target genes. Furthermore, around 5% to 10% of the target genes called from nearest TSS are not in the same TADs as the query variants, which implies the importance of considering the boundaries of TADs when predicting target genes. Most of the target genes called from eQTLs are not the nearest genes to the query variants. In summary, this comparison emphasizes the importance of incorporating genomic information other than the linear distance on chromosomes to assign target genes.

5.4.2 Target gene assignment provides functional hypothesis on ASB SNVs

The target genes assigned to regulatory variants can help explain the underlying mechanisms on how they involve in relevant pathways. For example, rs7438135 is an

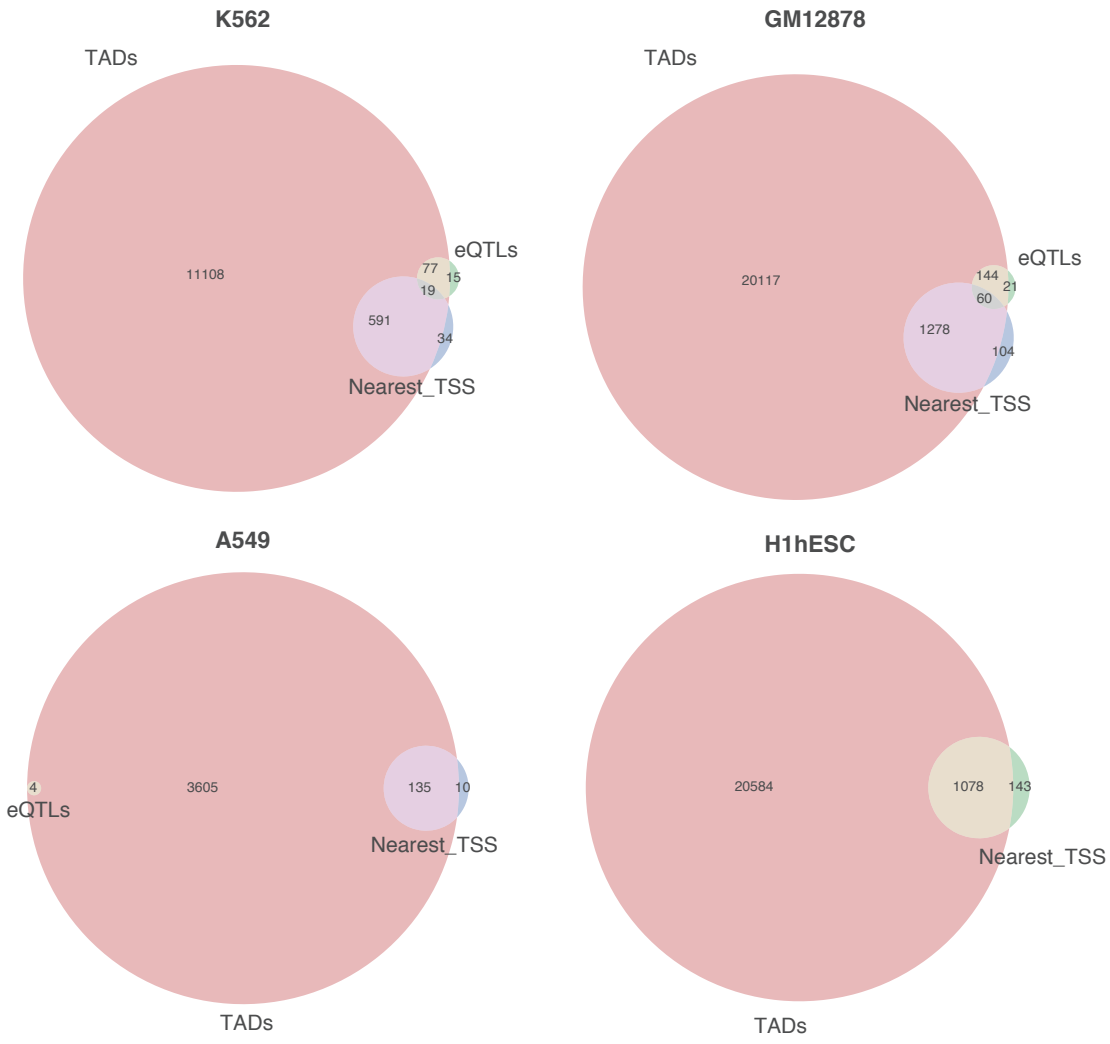


Figure 5.2: Comparison of target genes assigned from three approaches (Nearest TSS, TADs, and eQTLs) on ASB SNVs. Only the four cell lines with available TADs information are shown here.

ASB SNV called in A549, a lung cancer cell line. rs7438135 shows different binding affinity on two alleles in a TF ChIP-seq experiment targeting FOSL2 treated with dexamethasone. Dexamethasone is widely used in cancer patients' treatment to suppress the growth of non-small cell lung cancer. However, the underlying mechanisms remain unclear. Previous studies found dexamethasone is likely to induce apoptosis of A549 cells via the TGF-β1/Smad2 pathway [131], which is regulated by FOSL2 [132].

UGT2B4 was assigned to rs7438135 as the target gene by incorporating evidence from both TADs and eQTLs (Figure 5.3). The UGTs (UDP-glucuronosyltransferase enzymes) were found to affect cancer progression and drug resistance [133], including UGT2 as one of the four families of UGT. UGT enzymes are highly expressed in tissues such as liver and intestine, but the UGT subfamily members also have tissue-specific expression in many other tissues, including UGT2B4 in lung [134]. The regulation of UGTs involves multiple signal pathways and TFs across various tissues. Therefore, the potential regulation between FOSL2 and UGT2B4 found in this case can help explain how the regulatory variant rs7438135 might be functional in regulating UGT enzyme expression in lung cancer cells, which might be further involved in cancer progression and drug resistance.

5.4.3 Tissue-specific functions of ASB SNVs

Some of the target genes assigned to ASB SNVs have tissue-specific expression in corresponding cell lines, such as CD37 in K562 and UGT2B4 in A549. It indicates that the corresponding ASB SNVs can have tissue-specific regulatory functions. A previous study on DNase footprints shows that tissue-specific activity spectra of open chromatin regions is negatively correlated with the frequency of allele imbalance for the heterozygous sites within regions [135]. In other words, this could indicate that the sites with allele-specific regulatory patterns are more likely to have tissue-specific functions. To test if this is true for ASB SNVs, the ratio of target genes with tissue-specific expression was compared between ASB and non-ASB SNVs (Figure 5.4). However, similar ratios were found for SNVs in GM12878 and K562. The ASB SNVs in A549 show a higher ratio of tissue-specific genes than non-ASB SNVs, but only four target genes were assigned in this case. The results here are limited by the number of assigned target genes in each cell line. Also, the ASB SNVs were called

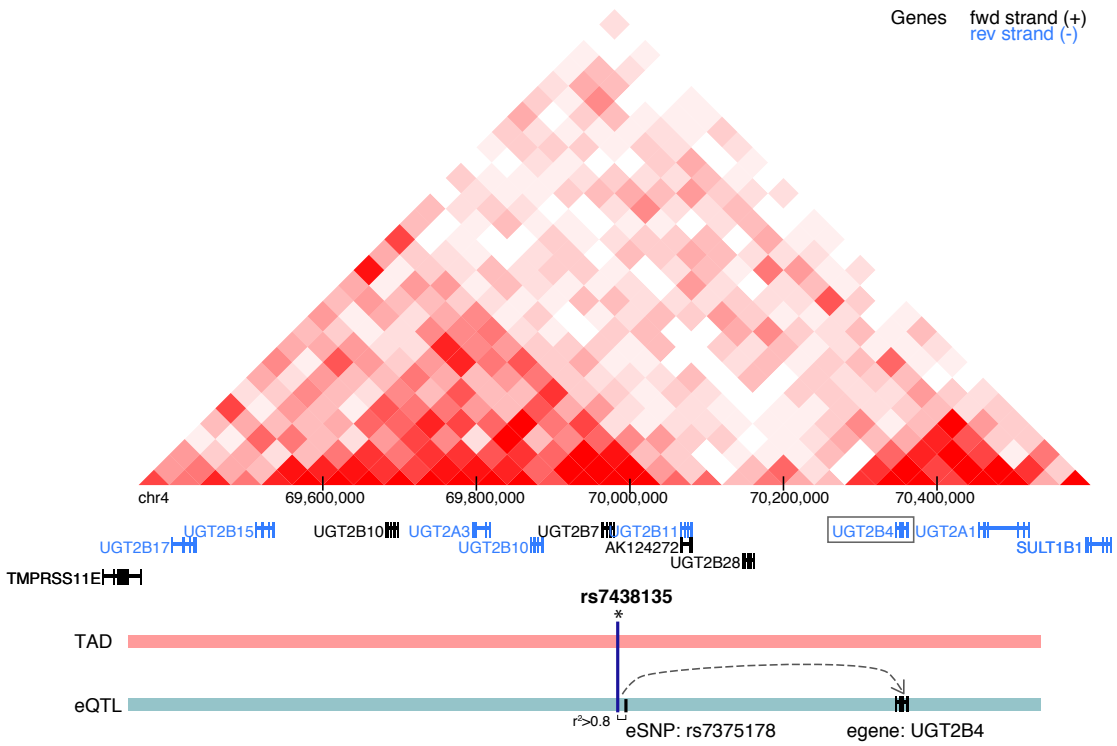


Figure 5.3: An example of target gene assignment on an ASB SNV by incorporating TADs and eQTLs evidence. The ASB SNV (rs7438135) called from A549 cell line is predicted to regulate a downstream gene UGT2B4. This gene in the same TAD from A549. The egene from eQTL studies in lung tissue is associated with an eSNP (rs7375178), which is in strong LD with rs7438135.

with a stringent threshold for extreme imbalance ratio between two alleles. Further studies on ASB SNVs involving a broader spectrum of imbalance ratios might lead to more solid conclusions.

5.5 Discussion

This study presented a workflow to incorporate tissue-specific genomic information from Hi-C and eQTL studies for target gene assignment. While the analysis here was performed on ASB SNVs, it can be generally applied for other putative regulatory variants, such as variants from GWAS studies, to understand their tissue-specific functions.

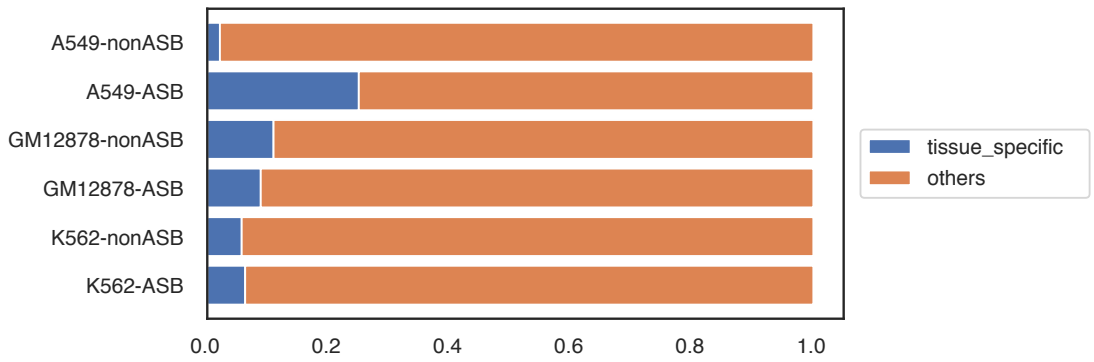


Figure 5.4: The ratio of genes with tissue-specific functions over target genes of ASB SNVs in three cell lines. The target genes were assigned to ASB SNVs by incorporating information from TADs and eQTLs.

The TADs regions used in this study have relatively low resolutions (∼40kb). With more Hi-C datasets in higher resolutions being generated, the TADs with finer structures can be defined. The 4D Nucleome (4DN) project has released many datasets from 3D conformation assays across various tissues and cell types. Efforts are needed to coordinate TADs annotations from different experiments and apply them in the corresponding tissue context. Chromatin loops identified from Hi-C and ChIA-PET can also be used to assign target genes, assuming that the regulatory variant and target gene are within the two ends of the loop structures. Moreover, as a higher-order chromatin structure, the compartments can also be useful information for target gene assignment. The two compartments with distinct gene transcription activities were identified previously [33], while more recent studies combining annotations from histone modifications further partitioned them into six smaller sub-compartments.

The target genes were assigned from eQTL studies with a threshold of $R^2 > 0.8$, which can be overly stringent. A looser threshold of 0.6 was tried, which includes 20% more variants with assigned egenes after LD expansion. However, since the measure-

ment of R^2 is largely dependent on allele frequency of the paired variants, using a strict threshold on R^2 might not be the most appropriate approach. Studies have been performed to partition the whole genome into LD blocks based on patterns in R^2 or D' values [136]. Assigning target genes in the same LD block with the query variant can be a more robust way to incorporate the LD structure. The variety of LD structure across populations also needs to be considered when applying to individual samples other than the cell lines in this study. Furthermore, QTLs from other quantitative traits relevant to gene regulatory activity can be incorporated with the same scheme as in eQTLs, such as the caQTLs associated with chromatin associability measured from ATAC-seq experiments [48]. Since caQTL data are currently available in limited cell types (lymphoblastoid and T cells), here I propose a computational method to leverage knowledge from caQTLs and eQTLs in different tissues. I borrowed the idea from TWAS that tissue-specific gene expression profile can be imputed based on genotypes [137] (Figure 5.5). Thus, the expression profile of individuals from caQTL studies can be imputed in various tissues with available eQTL datasets, for example the 49 human tissues from GTEx project. The correlation analysis between gene expression levels and ATAC-seq reads can be performed to identify the tissue-specific regulatory variants within ATAC-seq peaks.

On the other hand, some of the ASB SNVs in this study might have tissue-specific functions involved in specific biological pathways. Comparisons of target gene expression between two alleles could help further understand their functional consequences. This comparison will need a phased genome with known genotypes of variants on each allele, which is available for cell lines such as GM12878.

In summary, more efforts in computational approaches are needed to incorporate genomic information involving the coordination of data from various resources. As-

signing target genes to regulatory variants in a tissue-specific context can contribute the understand their functions in specific biological pathways and further explain their roles in disease progression.

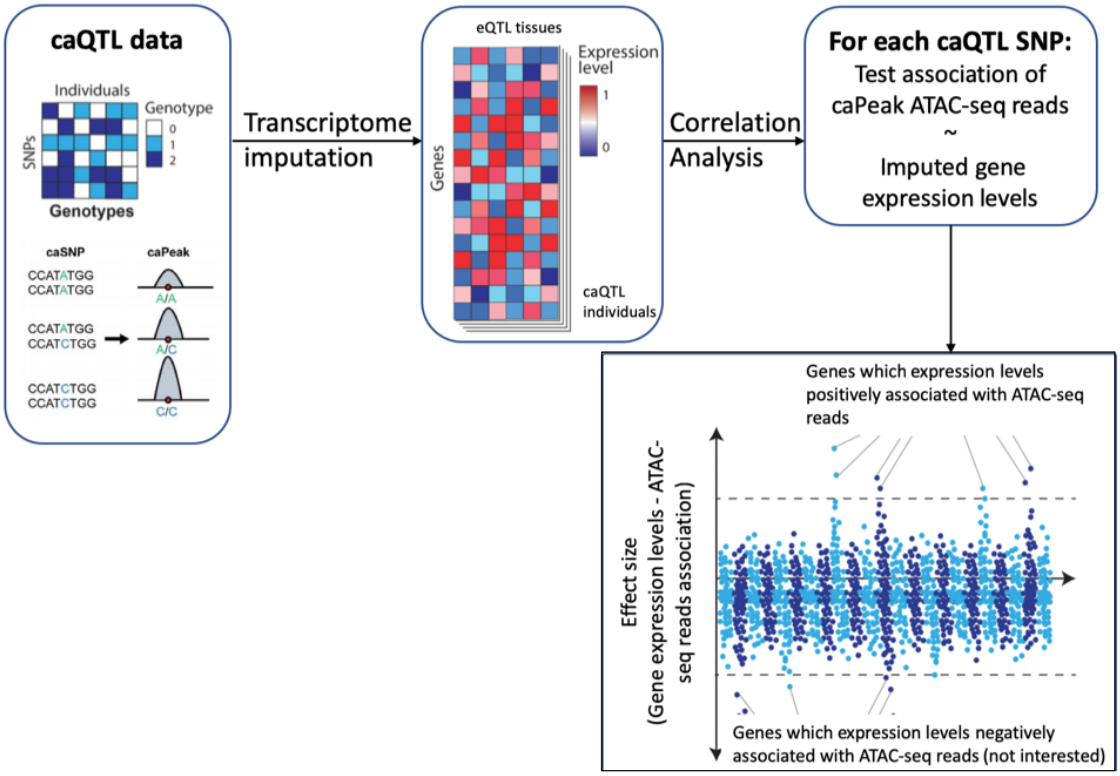


Figure 5.5: A proposed computational method to leverage knowledge from caQTLs and eQTLs in different tissues to identify tissue-specific regulatory variants within ATAC-seq peaks. The tissue-specific expression profile in eQTL tissues is imputed for caQTL individuals, following by correlation analysis between caPeak ATAC-seq reads and imputed gene expression level in each tissue. Figure adapted from [137].

CHAPTER VI

Conclusions and Future Directions

6.1 Summary

The main aim of my dissertation was to apply novel computational approaches to predict regulatory variants and provide functional hypotheses in the non-coding regions of human genome. I first developed a computational tool to predict the regulatory variants in 14 disease-associated enhancer and promoter regions, which achieved the best performance in CAGI5 challenge. I further extended this tool to predict tissue-specific functions of genome-wide regulatory variants. I calculated the prediction scores over 51 organs on approximately 80 million variants, which will be invaluable for future studies. I also designed the main updates of an online tool RegulomeDB to provide a user-friendly platform for quick annotation on non-coding variants. Finally, I explored a pipeline to assign target genes to putative regulatory variants with tissue-specific information from eQTL studies and Hi-C experiments.

In Chapter 2, I developed a computational tool, SURF, in the CAGI5 challenge. The challenge provided an unbiased comparison of computational tools for predicting regulatory variants with experimentally validated variants. Variants tested from MPRA experiments were used, including 17,500 SNVs in 9 promoters and 5 enhancers relevant to diseases. The 'true set' of regulatory variants were those show significant

effects on reporter gene expression levels from the MPRA datasets.

I trained random forest models using features including those binary ones retrieved from RegulomeDB old ranking scores, as well as additional numerical scores from ChIP-seq signals, PWMs, and prediction scores from DeepSEA using deep learning methods. In this way, the empirical tree from RegulomeDB ranking scores was replaced with sets of decision trees, including expanded features. The new probabilistic scores from SURF showed better performance in predicting regulatory variants compared to other participant groups. We found the binary features from direct functional genomics annotations provided complementary information compared to the transfer learning features from the DeepSEA model. This fact suggests the importance of incorporating direct annotations, especially for the cell types underrepresented in deep learning models.

One of the limitations in this study comes from the evaluation method, which mainly depended on AUROC. Adding a comparison on AUPR and correlation could improve the evaluation. Furthermore, although thousands of variants were included, they were limited by the variety of genomic regions. Thus, the models trained here might be biased to the genomics features on those specific regions and not be applicable for predicting genome-wide regulatory variants.

To extend the prediction of SURF, I further developed TURF in chapter 3. The extension was mainly on two aspects: 1) TURF provides a more robust prediction on a genome-wide scale; 2) TURF offers predictions of regulatory variants in a tissue-specific context. The first extension was achieved by training on regulatory variants covering a wider range of genomic regions. I called ASB SNVs in 6 cell lines from more than 600 ChIP-seq datasets, which are the putative regulatory variants. The mapping bias from aligning to the reference genome was corrected by building

personal genomes. Also, the copy number variation or aneuploidy regions were eliminated from the analysis. A total of 7,530 ASB SNVs were identified. In addition to training, those ASB SNVs can also lead to the following analysis to understand their regulatory functions and potential relationship to specific diseases or traits.

The second extension was relied on adding tissue-specific genomic features to the prediction models. The histone marks, the DHS, and DNase footprints were the most useful features to predict tissue-specific functions of regulatory variants. The final TURF tissue-specific scores leverage the regulatory patterns from other tissues by multiplying with TURF generic scores, while retaining the uniqueness of individual tissues by the tissue-specific features. I built an ensemble model to compensate for the variation from different cell lines in the training set. I also leverage the tissue-specific functional genomics annotations of tissues from the same organ to generate prediction scores covering most organs in ENCODE project.

Both the generic and tissue-specific scores from TURF presented overall better performance comparing to other state-of-the-art computational tools. I calculated TURF organ-specific prediction scores over 51 organs on the SNVs identified from GWAS studies and 1000 genome project, including approximately 80 million SNVs in total. I showed the ability to use TURF organ-specific scores to pick out the regulatory variants, which were validated to have organ-specific functions from previous studies. Moreover, many of the traits from GWAS Catalog displayed enrichment of organ-specific regulatory variants over the GWAS variants in their relevant organs. This enrichment indicates that the putative regulatory variants prioritized by TURF scores in the trait-relevant organs are likely to be involved in the corresponding traits. Following functional analysis on those regulatory variants might reveal underlying mechanisms for traits that are less-studied. Despite the GWAS variants, TURF can

be applied to any candidate list of variants on the whole genome and prioritize those with tissue/organ-specific functions. Furthermore, the TURF pipeline directly retrieves the functional genomics annotation and the organ terms from the ENCODE data portal. It will be easy to add new annotations or organs to the pipeline as more datasets are being released from ENCODE.

Due to the limited number of MPRA datasets, the performance from other computational tools was compared to TURF in only three cell lines. A more thorough comparison will be preferred with more MRRA datasets being available. In addition, the organ groups need to be refined to generate more intuitive annotations. More sophisticated normalization methods can be explored to correct the data availability imbalance across organs and apply to the background set for queries in different scenarios.

In chapter 4, I designed some main updates to RegulomeDB, which is an online tool for annotating non-coding variants with functional genomics evidence and prediction scores. We updated the RegulomeDB web server to display the query results with interactive charts grouped into six categories. These charts shown on the initial results page give users a quick and clear summary of the functional genomics annotations from different experiments to further explore their functions. We also expanded the data source in RegulomeDB database to include ENCODE phase 3 data and eQTLs from GTEx. The pipeline to generate prediction scores from SURF was also incorporated. The SURF scores are shown on the results page with the original ranking scores.

Future updates on RegulomeDB include integrating TURF scores into the system. The results page will have the option only to show tissue/organ-specific functional genomics annotations. Moreover, the data source for annotation will be further

expanded to include TADs from Hi-C experiments. Thus, users will have more references to make functional hypotheses on their query variants.

Finally, I explored the pipeline to assign target genes to ASB SNVs called from chapter 3 to help explain their functions and associations to disease progression. I compared the target genes assigned from three approaches, including the nearest genes on linear distance, the genes within the same TAD, and the genes from eQTL studies associated with an eSNP. The assignment was done in a tissue-specific manner. I found that the 'nearest genes' often have small overlapping with the genes from TADs or eQTLs evidence, emphasizing the need to incorporate tissue-specific genomic information from chromatin structure and association studies on gene expression levels. Using those two approaches, I presented an example of the target gene (UGT2B4) assigned to an ASB SNV (rs7438135) from lung cancer cells with the treatment of dexamethasone. Previous studies have shown that the expression of UGT2B4 is associated with lung cancer progression and drug resistance, which relies on the regulation from TFs in a lung-specific context. Further functional studies on the regulatory relationship between rs7438135 and UGT2B4 might help explain the underlying mechanism in lung cancer.

While the ASB SNVs were used as an example set, this workflow can be applied to other putative variants. Pipelines on incorporating TADs from Hi-C experiments covering more tissues are needed to annotate variants in a more general context. The LD structure can also be incorporated into the annotation from eQTLs in a more robust way other than a strict threshold on R^2 values.

6.2 Future directions

6.2.1 Refining TURF prediction algorithm

The feature set used in TURF generic scores can be categorized into four groups: 1) Binary features of direct annotations from functional genomics assays. 2) Quantiles describing the ChIP-seq signal profiles over various biosamples. 3) PWM matching and information content change capturing the sequence information on TF motifs. 4) Prediction score from DeepSEA trained from deep learning techniques. New features from each of the four groups can be explored.

1) Annotations from ATAC-seq might provide complementary information compared to DNase-seq.

2) The quantiles are currently only calculated from 200 ChIP-seq experiments from ENCODE phases 1 and 2 due to computation time limitations. It is unclear if adding more experiments will improve the prediction performance or the 200 experiments are already representative to describe the ChIP-seq profiles. It is also possible to add quantiles of DNase-seq signals to the feature set. Moreover, the signals were extracted from each position on the genome in base resolution. Some other tools used 'valley scores' to describe the signal pattern over nearby bases, which could be more informative than the signal in a single base. Also, the interpretation of this feature set needs to be explored more.

3) The main limitation of this feature set is that the PWM matching regions were aligned to the reference genome. While this captures most TF motifs, it is likely to miss the regions where a variant on the alternative allele creates a new TF binding site. This limitation can be solved using a looser threshold on selecting the kmers matching to each PWM or doing an additional mapping step with the alternative genotype of each query variant, which can be time-consuming. Furthermore, PWM

has been found not the best way to represent the TF binding affinity for each position on the motif. The SNP effect matrix from a recent paper that predicts TF binding affinity changes on SNPs can be used as a substitute for the PWM features [21].

4) DeepSEA scores were used as a typical model using deep learning methods. The advantage is that the regulatory grammars in various cell types from different assays were learned simultaneously. The functional consequence of each query variant is represented by a functional score combining all those regulatory grammars. However, the regulatory grammar patterns based on DNA sequences might be failing to capture the regulatory network in all cell types, especially for those underrepresented ones. In TURF, the direct annotations help detect features in less-studied cell types and interpret the model, while the scores from deep learning capture more consensus regulatory grammars involving a massive number of assays. It is possible to retrain the DeepSEA score with more experiments and make it more suitable for the TURF algorithm.

On the other hand, annotations from histone mark ChIP-seq and DNase-seq assays were used in the tissue-specific part, selected based on data availability and importance in feature ranking. More histone marks can be incorporated in the future as data being available. In addition, although the chromatin structure information in this dissertation was used in assigning target genes, it is possible to incorporate this tissue-specific feature set. The ratio of variants overlapping TADs was compared between positive versus negative training sets. However, no significant difference was found. The comparison was limited by the resolution and number of Hi-C experiments. Finer structures from TADs with histone mark information might be able to distinguish ASB and non-ASB SNVs, which can be used as additional features to predict tissue-specific functions.

Furthermore, it is also possible to incorporate validated variants from assays other than ChIP-seq into the training step to refine the TURF prediction algorithm. For example, the assays directly measuring regulatory activities, including MPRA and STAR-seq, can be incorporated. A framework that learns complementary information from different assays will need to be explored.

6.2.2 Incorporating TURF scores and target gene assignment to RegulomeDB

Other than the updates on data resources, we plan to add two new features to the RegulomeDB web server. The first is to incorporate TURF organ-specific scores. It is relatively straightforward since the organ terms can be retrieved from the JSON file currently associated with each experiment. In the initial results page for query variant, the interactive charts on underlying functional genomic evidence will be able to show specific tissues/organs. The heatmap of TURF scores for all query variants across 51 organs (as shown in Figure 3.8) will also be shown on the results page, which will be useful for the study of regulatory variants in specific pathways or diseases. The normalization of organ-specific scores was done by comparing them to the background set from all GWAS variants after LD expansion. However, the normalized scores relative to the query set from user input might also be useful to show.

The target gene assignment was performed on SNV variants in chapter 5, but we plan to perform this analysis for all query variants through RegulomeDB uniformly. Some challenges need to be addressed. The first is to integrate TADs regions from Hi-C experiments. These data are mainly available from the 4DN project, but additional processing steps will be required for those experiments not already called TADs regions. In addition, annotations need to be organized in a tissue-specific manner, which will involve manual curation to map to the relevant tissues in RegulomeDB.

This manual curation is also necessary to integrate eQTLs from GTEx tissues. More broadly, it is often challenging to incorporate the functional genomics information from different projects, which involves various cell lines and human tissues. Ontology analysis could help map between those samples.

6.2.3 Extending prediction to other genetic variations

While this dissertation focuses on SNVs, other genetic variations also play roles in regulatory functions. Perhaps it is most straightforward to extend TURF prediction to the variation of short insertions and deletions (indels). The simplest way might be using the current random forest model in the TURF algorithm. The binary features can still be generated by overlapping the query indel position to corresponding functional genomics evidence. In contrast, the numerical features can be obtained by taking the average or maximum over the values on all query indel positions. However, this scheme might be biased on the indels' length, especially for the binary features. Meanwhile, a different training set including indels might be needed to build a computational tool more adjusted to predicting regulatory indels.

6.2.4 ASB SNVs

The comparison of features in ASB and non-ASB SNVs has revealed some interesting patterns. For example, the ASB SNVs have a higher ratio of overlapping repressive histone marks relative to the non-ASB SNVs. One hypothesis is that the two alleles overlapping some of the ASB SNVs have different chromatin states. One allele is associated with active histone marks and has TF binding on the ASB SNV, while the other allele is associated with repressive histone marks and depleted on TF bindings. I tried to verify this hypothesis by performing allele-specific mapping on histone mark ChIP-seq reads but did not obtain definite conclusions, partly

due to the low coverage of some experiments. However, other functional studies on the ASB SNVs overlapping repressive histone marks might uncover the underlying mechanisms on how the allele-specific chromatin states are formed.

On the other hand, the hypothesis on the ASB SNVs are more enriched with tissue-specific functions was tested in chapter 5. This hypothesis is based on a previous observation that the tissue-specific activity spectra of DHS sites are negatively correlated with the frequency of allele imbalance for the heterozygous sites within the DHS [135]. However, the analysis in chapter 5 was limited by the number of target genes and the spectrum of allele imbalance ratio involved in the test. Further studies that overcome those two limitations might reveal new findings on the function of ASB SNVs.

6.3 Concluding remarks

In this dissertation, I developed computational tools to predict genome-wide regulatory variants in the non-coding regions of human genome. I developed machine learning models based on the scoring scheme from RegulomeDB v1.1, but significantly extended the annotation scale and improved prediction performance. The TURF algorithm (chapter 3) is able to provide tissue-specific prediction scores on a variety of tissues, which can be widely used to prioritize regulatory variants from association studies, such as GWAS. In the future, we plan to incorporate TURF into the RegulomeDB web server, as well as the pipeline in assigning target genes to putative regulatory variants (chapter 5). We hope this platform can have broad applications to guide researches on both known and de novo non-coding variants in various contexts including the precision treatment from clinical-relevant variants.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] E. H. Davidson, *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Elsevier, July 2010.

[2] K. Musunuru, A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt, K. V. Sachs, X. Li, H. Li, N. Kuperwasser, V. M. Ruda, J. P. Pirruccello, B. Muchmore, L. Prokunina-Olsson, J. L. Hall, E. E. Schadt, C. R. Morales, S. Lund-Katz, M. C. Phillips, J. Wong, W. Cantley, T. Racie, K. G. Ejebe, M. Orho-Melander, O. Melander, V. Koteliansky, K. Fitzgerald, R. M. Krauss, C. A. Cowan, S. Kathiresan, and D. J. Rader, "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus," *Nature*, vol. 466, pp. 714–719, Aug. 2010.

[3] J. R. Ecker, W. A. Bickmore, I. Barroso, J. K. Pritchard, Y. Gilad, and E. Segal, "Genomics: ENCODE explained," *Nature*, vol. 489, pp. 52–55, Sept. 2012.

[4] C. Wu, Y. C. Wong, and S. C. Elgin, "The chromatin structure of specific genes: II. disruption of chromatin structure during gene activity," *Cell*, vol. 16, pp. 807–814, Apr. 1979.

[5] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, pp. 311–322, Jan. 2008.

[6] L. Song and G. E. Crawford, "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," *Cold Spring Harb. Protoc.*, vol. 2010, p. db.prot5384, Feb. 2010.

[7] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nat. Methods*, vol. 10, pp. 1213–1218, Dec. 2013.

[8] D. J. Galas and A. Schmitz, "DNAse footprinting: a simple method for the detection of protein-DNA binding specificity," *Nucleic Acids Res.*, vol. 5, pp. 3157–3170, Sept. 1978.

[9] N. Ouyang and A. P. Boyle, "TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence," *Genome Res.*, vol. 30, pp. 1040–1046, July 2020.

[10] J. F. Degner, A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard, "DNase I sensitivity QTLs are a major determinant of human expression variation," *Nature*, vol. 482, pp. 390–394, Feb. 2012.

[11] B. Quach and T. S. Furey, "DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter," *Bioinformatics*, vol. 33, pp. 956–963, Apr. 2017.

[12] J. Kähärä and H. Lähdesmäki, "BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data," *Bioinformatics*, vol. 31, pp. 2852–2859, Sept. 2015.

[13] J. Piper, M. C. Elze, P. Cauchy, P. N. Cockerill, C. Bonifer, and S. Ott, "Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data," *Nucleic Acids Res.*, vol. 41, p. e201, Nov. 2013.

[14] R. I. Sherwood, T. Hashimoto, C. W. O'Donnell, S. Lewis, A. A. Barkal, J. P. van Hoff, V. Karun, T. Jaakkola, and D. K. Gifford, "Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape," *Nat. Biotechnol.*, vol. 32, pp. 171–178, Feb. 2014.

[15] E. G. Gusmao, C. Dieterich, M. Zenke, and I. G. Costa, "Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications," *Bioinformatics*, vol. 30, pp. 3143–3151, Nov. 2014.

[16] S. M. Lloyd and X. Bao, "Pinpointing the genomic localizations of Chromatin-Associated proteins: The yesterday, today, and tomorrow of ChIP-seq," *Curr. Protoc. Cell Biol.*, vol. 84, p. e89, Sept. 2019.

[17] B. L. Kidder, G. Hu, and K. Zhao, "ChIP-Seq: technical considerations for obtaining high-quality data," *Nat. Immunol.*, vol. 12, pp. 918–922, Sept. 2011.

[18] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'perceptron' algorithm to distinguish translational initiation sites in e. coli," *Nucleic Acids Res.*, vol. 10, pp. 2997–3011, May 1982.

[19] C. Tuerk and L. Gold, "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase," *Science*, vol. 249, pp. 505–510, Aug. 1990.

[20] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, W. Santana-Garcia, G. Tan, J. Chèneby, B. Ballester, F. Parcy, A. Sandelin, B. Lenhard, W. W. Wasserman, and A. Mathelier, "JASPAR 2020: update of the open-access database of transcription factor binding profiles," *Nucleic Acids Res.*, vol. 48, pp. D87–D92, Jan. 2020.

[21] S. S. Nishizaki, N. Ng, S. Dong, R. S. Porter, C. Morterud, C. Williams, C. Asman, J. A. Switzenberg, and A. P. Boyle, "Predicting the effects of SNPs on transcription factor binding affinity," *Bioinformatics*, vol. 36, pp. 364–372, Jan. 2020.

[22] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, pp. 317–330, Feb. 2015.

[23] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation," *Nat. Methods*, vol. 9, pp. 473–476, Mar. 2012.

[24] R. C. W. Chan, M. W. Libbrecht, E. G. Roberts, J. A. Bilmes, W. S. Noble, and M. M. Hoffman, "Segway 2.0: Gaussian mixture models and minibatch training," *Bioinformatics*, vol. 34, pp. 669–671, Feb. 2018.

[25] P. Kheradpour, J. Ernst, A. Melnikov, P. Rogov, L. Wang, X. Zhang, J. Alston, T. S. Mikkelsen, and M. Kellis, "Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay," *Genome Res.*, vol. 23, pp. 800–811, May 2013.

[26] R. Tewhey, D. Kotliar, D. S. Park, B. Liu, S. Winnicki, S. K. Reilly, K. G. Andersen, T. S. Mikkelsen, E. S. Lander, S. F. Schaffner, and P. C. Sabeti, "Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay," *Cell*, vol. 172, pp. 1132–1134, Feb. 2018.

[27] D. Shigaki, O. Adato, A. N. Adhikari, S. Dong, A. Hawkins-Hooker, F. Inoue, T. Juven-Gershon, H. Kenlay, B. Martin, A. Patra, D. D. Penzar, M. Schubach, C. Xiong, Z. Yan, A. P. Boyle, A. Kreimer, I. V. Kulakovskiy, J. Reid, R. Unger, N. Yosef, J. Shendure, N. Ahituv, M. Kircher, and M. A. Beer, "Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay," *Hum. Mutat.*, vol. 40, pp. 1280–1291, Sept. 2019.

[28] F. Inoue and N. Ahituv, "Decoding enhancers using massively parallel reporter assays," *Genomics*, vol. 106, pp. 159–164, Sept. 2015.

[29] R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S.-I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, and J. Shendure, "Massively parallel functional dissection of mammalian enhancers in vivo," *Nat. Biotechnol.*, vol. 30, pp. 265–270, Feb. 2012.

[30] W. Deng, J. W. Rupon, I. Krivega, L. Breda, I. Motta, K. S. Jahn, A. Reik, P. D. Gregory, S. Rivella, A. Dean, and G. A. Blobel, "Reactivation of developmentally silenced globin genes by forced chromatin looping," *Cell*, vol. 158, pp. 849–860, Aug. 2014.

[31] Z. Tang, O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L. M. Sachs, X. Ruan, C.-L. Wei, E. T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li, and Y. Ruan, "CTCF-Mediated human 3D genome architecture reveals chromatin topology for transcription," *Cell*, vol. 163, pp. 1611–1627, Dec. 2015.

[32] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing chromosome conformation," *Science*, vol. 295, pp. 1306–1311, Feb. 2002.

[33] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, pp. 289–293, Oct. 2009.

[34] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, pp. 1665–1680, Dec. 2014.

[35] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Y. Chew, P. Y. H. Huang, W.-J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. S. A. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. M. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W.-K.

Sung, E. T. Liu, C.-L. Wei, E. Cheung, and Y. Ruan, "An oestrogen-receptor-α-bound human chromatin interactome," *Nature*, vol. 462, pp. 58–64, Nov. 2009.

[36] J. H. Gibcus and J. Dekker, "The hierarchy of the 3D genome," *Mol. Cell*, vol. 49, pp. 773–782, Mar. 2013.

[37] J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, S. Zhong, and 4D Nucleome Network, "The 4D nucleome project," *Nature*, vol. 549, pp. 219–226, Sept. 2017.

[38] Z. Liang, G. Li, Z. Wang, M. N. Djekidel, Y. Li, M.-P. Qian, M. Q. Zhang, and Y. Chen, "BL-Hi-C is an efficient and sensitive approach for capturing structural and regulatory chromatin interactions," *Nat. Commun.*, vol. 8, Dec. 2017.

[39] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure, "Massively multiplex single-cell Hi-C," *Nat. Methods*, vol. 14, pp. 263–266, Mar. 2017.

[40] ENCODE Project Consortium, J. E. Moore, M. J. Purcaro, H. E. Pratt, C. B. Epstein, N. Shoresh, J. Adrian, T. Kawli, C. A. Davis, A. Dobin, R. Kaul, J. Halow, E. L. Van Nostrand, P. Freese, D. U. Gorkin, Y. Shen, Y. He, M. Mackiewicz, F. Pauli-Behn, B. A. Williams, A. Mortazavi, C. A. Keller, X.-O. Zhang, S. I. Elhajjajy, J. Huey, D. E. Dickel, V. Snetkova, X. Wei, X. Wang, J. C. Rivera-Mulia, J. Rozowsky, J. Zhang, S. B. Chhetri, J. Zhang, A. Victorsen, K. P. White, A. Visel, G. W. Yeo, C. B. Burge, E. Lécuyer, D. M. Gilbert, J. Dekker, J. Rinn, E. M. Mendenhall, J. R. Ecker, M. Kellis, R. J. Klein, W. S. Noble, A. Kundaje, R. Guigó, P. J. Farnham, J. M. Cherry, R. M. Myers, B. Ren, B. R. Graveley, M. B. Gerstein, L. A. Pennacchio, M. P. Snyder, B. E. Bernstein, B. Wold, R. C. Hardison, T. R. Gingeras, J. A. Stamatoyannopoulos, and Z. Weng, "Expanded encyclopaedias of DNA elements in the human and mouse genomes," *Nature*, vol. 583, pp. 699–710, July 2020.

[41] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson, "The NIH roadmap epigenomics mapping consortium," *Nat. Biotechnol.*, vol. 28, pp. 1045–1048, Oct. 2010.

[42] W. A. Nelson and E. E. Crone, "Genetics and analysis of quantitative traits. michael lynch , bruce walsh," *Q. Rev. Biol.*, vol. 74, pp. 225–225, June 1999.

[43] F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, and E. Eskin, "Identifying causal variants at loci with multiple signals of association," *Genetics*, vol. 198, pp. 497–508, Oct. 2014.

[44] C. Benner, C. C. A. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti, and M. Pirinen, "FINEMAP: efficient variable selection using summary data from genome-wide association studies," *Bioinformatics*, vol. 32, pp. 1493–1501, May 2016.

[45] R. Mägi, M. Horikoshi, T. Sofer, A. Mahajan, H. Kitajima, N. Franceschini, M. I. McCarthy, A. P. Morris, and COGENT-Kidney Consortium, T2D-GENES Consortium, "Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution," *Hum. Mol. Genet.*, vol. 26, pp. 3639–3650, Sept. 2017.

[46] GTEx Consortium, "The Genotype-Tissue expression (GTEx) project," *Nat. Genet.*, vol. 45, pp. 580–585, June 2013.

[47] GTEx Consortium, "The GTEx consortium atlas of genetic regulatory effects across human tissues," *Science*, vol. 369, pp. 1318–1330, Sept. 2020.

[48] N. Kumasaka, A. J. Knights, and D. J. Gaffney, "High-resolution genetic mapping of putative causal interactions between regions of open chromatin," *Nat. Genet.*, vol. 51, pp. 128–137, Jan. 2019.

[49] A. Tehranchi, B. Hie, M. Dacre, I. Kaplow, K. Pettie, P. Combs, and H. B. Fraser, "Fine-mapping cis-regulatory variants in diverse human populations," *Elife*, vol. 8, Jan. 2019.

[50] J. Rozowsky, A. Abyzov, J. Wang, P. Alves, D. Raha, A. Harmanci, J. Leng, R. Bjornson, Y. Kong, N. Kitabayashi, N. Bhardwaj, M. Rubin, M. Snyder, and M. Gerstein, "AlleleSeq: analysis of allele-specific expression and binding in a network framework," *Mol. Syst. Biol.*, vol. 7, p. 522, Aug. 2011.

[51] S. D. Bailey, C. Virtanen, B. Haibe-Kains, and M. Lupien, "ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments," *Bioinformatics*, vol. 31, pp. 3057–3059, Sept. 2015.

[52] J. Chen, J. Rozowsky, T. R. Galeev, A. Harmanci, R. Kitchen, J. Bedford, A. Abyzov, Y. Kong, L. Regan, and M. Gerstein, "A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals," *Nat. Commun.*, vol. 7, p. 11101, Apr. 2016.

[53] M. Cavalli, G. Pan, H. Nord, O. Wallerman, E. Wallén Arzt, O. Berggren, I. Elvers, M.-L. Eloranta, L. Rönnblom, K. Lindblad Toh, and C. Wadelius, "Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression," *Hum. Genet.*, vol. 135, pp. 485–497, May 2016.

[54] I. de Santiago, W. Liu, K. Yuan, M. O'Reilly, C. S. R. Chilamakuri, B. A. J. Ponder, K. B. Meyer, and F. Markowetz, "BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes," *Genome Biol.*, vol. 18, p. 39, Feb. 2017.

[55] M. Cavalli, N. Baltzer, H. M. Umer, J. Grau, I. Lemnian, G. Pan, O. Wallerman, R. Spalinskas, P. Sahlén, I. Grosse, J. Komorowski, and C. Wadelius, "Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases," *Sci. Rep.*, vol. 9, p. 2695, Feb. 2019.

[56] J. Zou, F. Hormozdiari, B. Jew, S. E. Castel, T. Lappalainen, J. Ernst, J. H. Sul, and E. Eskin, "Leveraging allelic imbalance to refine fine-mapping for eQTL studies," *PLoS Genet.*, vol. 15, p. e1008481, Dec. 2019.

[57] M. J. Li, M. Li, Z. Liu, B. Yan, Z. Pan, D. Huang, Q. Liang, D. Ying, F. Xu, H. Yao, P. Wang, J.-P. A. Kocher, Z. Xia, P. C. Sham, J. S. Liu, and J. Wang, "cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes," *Genome Biol.*, vol. 18, p. 52, Mar. 2017.

[58] K. M. de Lange, L. Moutsianas, J. C. Lee, C. A. Lamb, Y. Luo, N. A. Kennedy, L. Jostins, D. L. Rice, J. Gutierrez-Achury, S.-G. Ji, G. Heap, E. R. Nimmo, C. Edwards, P. Henderson, C. Mowat, J. Sanderson, J. Satsangi, A. Simmons, D. C. Wilson, M. Tremelling, A. Hart, C. G. Mathew, W. G. Newman, M. Parkes, C. W. Lees, H. Uhlig, C. Hawkey, N. J. Prescott, T. Ahmad, J. C. Mansfield, C. A. Anderson, and J. C. Barrett, "Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease," *Nat. Genet.*, vol. 49, pp. 256–261, Feb. 2017.

[59] L. Yengo, J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, P. M. Visscher, and GIANT Consortium, "Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry," *Hum. Mol. Genet.*, vol. 27, pp. 3641–3649, Oct. 2018.

[60] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson, "The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," *Nucleic Acids Res.*, vol. 47, pp. D1005–D1012, Jan. 2019.

[61] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre, "Benefits and limitations of genome-wide association studies," *Nat. Rev. Genet.*, vol. 20, pp. 467–484, Aug. 2019.

[62] A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder, "Annotation of functional variation in personal genomes using RegulomeDB," *Genome Res.*, vol. 22, pp. 1790–1797, Sept. 2012.

[63] L. D. Ward and M. Kellis, "HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease," *Nucleic Acids Res.*, vol. 44, pp. D877–81, Jan. 2016.

[64] M. A. Beer, "Predicting enhancer activity and variant impact using gkm-SVM," *Hum. Mutat.*, vol. 38, pp. 1251–1258, Sept. 2017.

[65] H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, T. R. Gaunt, and C. Campbell, "An integrative approach to predicting the functional effects of non-coding and coding sequence variation," *Bioinformatics*, vol. 31, pp. 1536–1543, May 2015.

[66] G. R. S. Ritchie, I. Dunham, E. Zeggini, and P. Flicek, "Functional annotation of noncoding sequence variants," *Nat. Methods*, vol. 11, pp. 294–296, Mar. 2014.

[67] M. Krawczak, E. V. Ball, P. Stenson, and D. N. Cooper, "HGMD: The human gene mutation database," in *Bioinformatics: Databases and Systems*, pp. 99–104, Boston: Kluwer Academic Publishers, 2006.

[68] D. Backenroth, Z. He, K. Kiryluk, V. Boeva, L. Pethukova, E. Khurana, A. Christiano, J. D. Buxbaum, and I. Ionita-Laza, "FUN-LDA: A latent dirichlet allocation model for predicting Tissue-Specific functional effects of noncoding variation: Methods and applications," *Am. J. Hum. Genet.*, vol. 102, pp. 920–942, May 2018.

[69] Q. Lu, R. L. Powles, Q. Wang, B. J. He, and H. Zhao, "Integrative Tissue-Specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies," *PLoS Genet.*, vol. 12, p. e1005947, Apr. 2016.

[70] Z. He, L. Liu, K. Wang, and I. Ionita-Laza, "A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs," *Nat. Commun.*, vol. 9, p. 5199, Dec. 2018.

[71] D. Huang, X. Yi, S. Zhang, Z. Zheng, P. Wang, C. Xuan, P. C. Sham, J. Wang, and M. J. Li, "GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits," *Nucleic Acids Res.*, vol. 46, pp. W114–W120, July 2018.

[72] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nat. Methods*, vol. 12, pp. 931–934, Oct. 2015.

[73] J. Zhou, C. Y. Park, C. L. Theesfeld, A. K. Wong, Y. Yuan, C. Scheckel, J. J. Fak, J. Funk, K. Yao, Y. Tajima, A. Packer, R. B. Darnell, and O. G. Troyanskaya, "Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk," *Nat. Genet.*, vol. 51, pp. 973–980, June 2019.

[74] L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, and K. K.-H. Farh, "Predicting the clinical impact of human mutation with deep neural networks," *Nat. Genet.*, vol. 50, pp. 1161–1170, Aug. 2018.

[75] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," Apr. 2017.

[76] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," Feb. 2016.

[77] S. Zhang, Y. He, H. Liu, H. Zhai, D. Huang, X. Yi, X. Dong, Z. Wang, K. Zhao, Y. Zhou, J. Wang, H. Yao, H. Xu, Z. Yang, P. C. Sham, K. Chen, and M. J. Li, "regbase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants," *Nucleic Acids Res.*, vol. 47, p. e134, Dec. 2019.

[78] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, "Hi-C: a comprehensive technique to capture the conformation of genomes," *Methods*, vol. 58, pp. 268–276, Nov. 2012.

[79] J. A. Beagan and J. E. Phillips-Cremins, "On the existence and functionality of topologically associating domains," *Nat. Genet.*, vol. 52, pp. 8–16, Jan. 2020.

[80] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, Apr. 2012.

[81] E. P. Nora, A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, and B. G. Bruneau, "Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization," *Cell*, vol. 169, pp. 930–944.e22, May 2017.

[82] I. Jung, A. Schmitt, Y. Diao, A. J. Lee, T. Liu, D. Yang, C. Tan, J. Eom, M. Chan, S. Chee, Z. Chiang, C. Kim, E. Masliah, C. L. Barr, B. Li, S. Kuan, D. Kim, and B. Ren, "A compendium of promoter-centered long-range chromatin interactions in the human genome," *Nat. Genet.*, vol. 51, pp. 1442–1449, Oct. 2019.

[83] M. Song, X. Yang, X. Ren, L. Maliskova, B. Li, I. R. Jones, C. Wang, F. Jacob, K. Wu, M. Traglia, T. W. Tam, K. Jamieson, S.-Y. Lu, G.-L. Ming, Y. Li, J. Yao, L. A. Weiss, J. R. Dixon, L. M. Judge, B. R. Conklin, H. Song, L. Gan, and Y. Shen, "Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes," *Nat. Genet.*, vol. 51, pp. 1252–1262, Aug. 2019.

[84] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, pp. D1001–D1006, Jan. 2014.

[85] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, pp. 9362–9367, June 2009.

[86] S. S. Nishizaki and A. P. Boyle, "Mining the unknown: Assigning function to noncoding single nucleotide polymorphisms," *Trends Genet.*, vol. 33, pp. 34–45, Jan. 2017.

[87] M. J. Li, L. Y. Wang, Z. Xia, P. C. Sham, and J. Wang, "GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications," *Nucleic Acids Res.*, vol. 41, pp. W150–8, July 2013.

[88] M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nat. Genet.*, vol. 46, pp. 310–315, Mar. 2014.

[89] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat. Biotechnol.*, vol. 33, pp. 831–838, Aug. 2015.

[90] D. Quang and X. Xie, "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences," *Nucleic Acids Res.*, vol. 44, p. e107, June 2016.

[91] M. Wang, C. Tai, W. E, and L. Wei, "DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants," *Nucleic Acids Res.*, vol. 46, p. e69, June 2018.

[92] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks," *Genome Res.*, vol. 28, pp. 739–750, May 2018.

[93] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender, "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Res.*, vol. 34, pp. D108–10, Jan. 2006.

[94] J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin, "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update," *Nucleic Acids Res.*, vol. 36, pp. D102–6, Jan. 2008.

[95] D. E. Newburger and M. L. Bulyk, "UniPROBE: an online database of protein binding microarray data on protein-DNA interactions," *Nucleic Acids Res.*, vol. 37, pp. D77–82, Jan. 2009.

[96] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale, "DNA-binding specificities of human transcription factors," *Cell*, vol. 152, pp. 327–339, Jan. 2013.

[97] A. P. Boyle, L. Song, B.-K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey, "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells," *Genome Res.*, vol. 21, pp. 456–464, Mar. 2011.

[98] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard, "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data," *Genome Res.*, vol. 21, pp. 447–455, Mar. 2011.

[99] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 1947–1958, Nov. 2003.

[100] A. Kreimer, Z. Yan, N. Ahituv, and N. Yosef, "Meta-analysis of massive parallel reporter assay enables functional regulatory elements prediction." Oct. 2017.

[101] C. Melton, J. A. Reuter, D. V. Spacek, and M. Snyder, "Recurrent somatic mutations in regulatory regions of human cancer genomes," *Nat. Genet.*, vol. 47, pp. 710–716, July 2015.

[102] A. Sharma, C. Jiang, and S. De, "Dissecting the sources of gene expression variation in a pan-cancer analysis identifies novel regulatory mutations," *Nucleic Acids Res.*, vol. 46, pp. 4370–4381, May 2018.

[103] S. Dong and A. P. Boyle, "Predicting functional variants in enhancer and promoter elements using RegulomeDB," *Hum. Mutat.*, vol. 40, pp. 1292–1298, Sept. 2019.

[104] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, and J. M. Cherry, "The encyclopedia of DNA elements (ENCODE): data portal update," *Nucleic Acids Res.*, vol. 46, pp. D794–D801, Jan. 2018.

[105] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, Sept. 2012.

[106] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, pp. 823–837, May 2007.

[107] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, pp. 1497–1502, June 2007.

[108] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, pp. 553–560, Aug. 2007.

[109] M. Kasowski, S. Kyriazopoulou-Panagiotopoulou, F. Grubert, J. B. Zaugg, A. Kundaje, Y. Liu, A. P. Boyle, Q. C. Zhang, F. Zakharia, D. V. Spacek, J. Li, D. Xie, A. Olarerin-George, L. M. Steinmetz, J. B. Hogenesch, M. Kellis, S. Batzoglou, and M. Snyder, "Extensive variation in chromatin states across humans," *Science*, vol. 342, pp. 750–752, Nov. 2013.

[110] 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, Oct. 2015.

[111] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nat. Genet.*, vol. 43, pp. 491–498, May 2011.

[112] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing," *Genome Res.*, vol. 21, pp. 974–984, June 2011.

[113] S. T. Sherry, "dbSNP: the NCBI database of genetic variation," 2001.

[114] B. Yang, W. Zhou, J. Jiao, J. B. Nielsen, M. R. Mathis, M. Heydarpour, G. Lettre, L. Folkersen, S. Prakash, C. Schurmann, L. Fritsche, G. A. Farnum, M. Lin, M. Othman, W. Hornsby, A. Driscoll, A. Levasseur, M. Thomas, L. Farhat, M.-P. Dubé, E. M. Isselbacher, A. Franco-Cereceda, D.-C. Guo, E. P. Bottinger, G. M. Deeb, A. Booher, S. Kheterpal, Y. E. Chen, H. M. Kang, J. Kitzman, H. J. Cordell, B. D. Keavney, J. A. Goodship, S. K. Ganesh, G. Abecasis, K. A. Eagle, A. P. Boyle, R. J. F. Loos, P. Eriksson, J.-C. Tardif, C. M. Brummett, D. M. Milewicz, S. C. Body, and C. J. Willer, "Protein-altering and regulatory genetic variants near GATA4 implicated in bicuspid aortic valve," *Nat. Commun.*, vol. 8, p. 15481, May 2017.

[115] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos, "Systematic localization of common disease-associated variation in regulatory DNA," *Science*, vol. 337, pp. 1190–1195, Sept. 2012.

[116] R. M. Glickman and P. H. Green, "The intestine as a source of apolipoprotein A1," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, pp. 2569–2573, June 1977.

[117] J. J. Diehn, M. Diehn, M. F. Marmor, and P. O. Brown, "Differential gene expression in anatomical compartments of the human eye," *Genome Biol.*, vol. 6, p. R74, Aug. 2005.

[118] Z. Wang, D. L. White, R. Hoogeveen, L. Chen, E. A. Whitsel, P. A. Richardson, S. S. Virani, J. M. Garcia, H. B. El-Serag, and L. Jiao, "Anti-Hypertensive medication use, soluble receptor for glycation end products and risk of pancreatic cancer in the women's health initiative study," *J. Clin. Med. Res.*, vol. 7, Aug. 2018.

[119] S. Dong and A. P. Boyle, "Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome." Mar. 2021.

[120] Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, M. Hardison, R. Person, M. R. Bekheirnia, M. S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S. E. Plon, J. R. Lupski, A. L. Beaudet, R. A. Gibbs, and C. M. Eng, "Clinical whole-exome sequencing for the diagnosis of mendelian disorders," *N. Engl. J. Med.*, vol. 369, pp. 1502–1511, Oct. 2013.

[121] E. A. Worthey, A. N. Mayer, G. D. Syverson, D. Helbling, B. B. Bonacci, B. Decker, J. M. Serpe, T. Dasu, M. R. Tschannen, R. L. Veith, M. J. Basehore, U. Broeckel, A. Tomita-Mitchell, M. J. Arca, J. T. Casper, D. A. Margolis, D. P. Bick, M. J. Hessner, J. M. Routes, J. W. Verbsky, H. J. Jacob, and D. P. Dimmock, "Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease," *Genet. Med.*, vol. 13, pp. 255–262, Mar. 2011.

[122] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad, "Exome sequencing identifies the cause of a mendelian disorder," *Nat. Genet.*, vol. 42, pp. 30–35, Jan. 2010.

[123] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *N. Engl. J. Med.*, vol. 372, pp. 793–795, Feb. 2015.

[124] D. Horgan, M. Romao, and R. Hastings, "Pulling the Strands Together: MEGA Steps to Drive European Genomics and Personalised Medicine," *Biomed Hub*, vol. 2, pp. 169–179, Nov. 2017.

[125] J. M. Gaziano, J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, P. Guarino, M. Aslan, D. Anderson, R. LaFleur, T. Hammond, K. Schaa, J. Moser, G. Huang, S. Muralidhar, R. Przygodzki, and T. J. O'Leary, "Million veteran program: A mega-biobank to study genetic influences on health and disease," *J. Clin. Epidemiol.*, vol. 70, pp. 214–223, Feb. 2016.

[126] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, p. R25, Mar. 2009.

[127] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization," *Nat. Methods*, vol. 9, pp. 215–216, Feb. 2012.

[128] L.-H. Chang, S. Ghosh, and D. Noordermeer, "TADs and their borders: Free movement or building a wall?," *J. Mol. Biol.*, vol. 432, pp. 643–652, Feb. 2020.

[129] Y. Wang, F. Song, B. Zhang, L. Zhang, J. Xu, D. Kuang, D. Li, M. N. K. Choudhary, Y. Li, M. Hu, R. Hardison, T. Wang, and F. Yue, "The 3D genome browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions," *Genome Biol.*, vol. 19, p. 151, Oct. 2018.

[130] A. R. Sonawane, J. Platig, M. Fagny, C.-Y. Chen, J. N. Paulson, C. M. Lopes-Ramos, D. L. DeMeo, J. Quackenbush, K. Glass, and M. L. Kuijjer, "Understanding Tissue-Specific gene regulation," *Cell Rep.*, vol. 21, pp. 1077–1088, Oct. 2017.

[131] X. Feng, H. Fei, and L. Hu, "Dexamethasone induced apoptosis of A549 cells via the TGF-β1/Smad2 pathway," *Oncol. Lett.*, Dec. 2017.

[132] J. Wang, D. Sun, Y. Wang, F. Ren, S. Pang, D. Wang, and S. Xu, "FOSL2 positively regulates TGF-β1 signalling in non-small cell lung cancer," *PLoS One*, vol. 9, p. e112150, Nov. 2014.

[133] E. P. Allain, M. Rouleau, E. Lévesque, and C. Guillemette, "Emerging roles for UDP-glucuronosyltransferases in drug resistance and cancer progression," *Br. J. Cancer*, vol. 122, pp. 1277–1287, Apr. 2020.

[134] R. Meech, D. G. Hu, R. A. McKinnon, S. N. Mubarokah, A. Z. Haines, P. C. Nair, A. Rowland, and P. I. Mackenzie, "The UDP-Glycosyltransferase (UGT) superfamily: New members, new functions, and novel paradigms," *Physiol. Rev.*, vol. 99, pp. 1153–1222, Apr. 2019.

[135] M. T. Maurano, E. Haugen, R. Sandstrom, J. Vierstra, A. Shafer, R. Kaul, and J. A. Stamatoyannopoulos, "Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo," *Nat. Genet.*, vol. 47, pp. 1393–1401, Dec. 2015.

[136] S. A. Kim, C.-S. Cho, S.-R. Kim, S. B. Bull, and Y. J. Yoo, "A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs," *Bioinformatics*, vol. 34, pp. 388–397, Feb. 2018.

[137] E. Cano-Gamez and G. Trynka, "From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases," *Front. Genet.*, vol. 11, p. 424, May 2020.