Decoding Regulatory Variants with Computational Methods in Non-coding Regions of the Human Genome

by

Nanxiang Zhao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2023

Doctoral Committee:

Associate Professor Alan Boyle, Chair
Professor Harm Derksen, Northeastern University
Associate Professor Jacob Kitzman
Professor Kayvan Najarian
Assistant Professor Joshua Welch

Nanxiang Zhao
samzhao@umich.edu
ORCID iD: 0000-0003-3124-0958

To my loving family and friends

ACKNOWLEDGEMENTS

My dissertation would not have been possible without the invaluable help, support, and love of many people. I would like to express my deepest gratitude to my mentor, Alan Boyle, for his support and encouragement. Alan has been a wonderful mentor, always believing in me, encouraging me to learn and explore, and treating me like family. When I first joined our computational lab, I had little experience with programming, but Alan's guidance and dedication to our growth helped me develop my skills. He was always available to meet with us individually, and he never turned down any of our ideas, always encouraging us to explore and experiment. Alan's curiosity and passion for science made learning a joy, and he taught me the most important aspect of scientific inquiry, which is how to ask the right questions. I cannot thank Alan enough for his mentorship and support, which have been instrumental in the success of my dissertation.

I would like to express my gratitude to all members of the Boyle lab family. It has been a great pleasure working and spending the past six years with all of you. Thank you for always taking care of one another and creating a supportive and familial environment within the lab. I would like to extend a special thanks to Adam Diehl for his invaluable technical assistance and for greatly helping me with my manuscripts. Jessica Switzenberg, thank you for your meticulous editing and for teaching me how to write better. Christopher Castro, thank you for your great advice on both work and life matters. I would also like to thank Ningxin

Ouyang, Bradley Crone, Camille Mumm, Torrin McDonald, Melissa Englund, Sierra Nishizaki, Andrea Valenzuela, Breanna McBean, and Kinsey Van Deynze for our stimulating discussions and shared struggles throughout my Ph.D. journey. Finally, I would like to express my gratitude to Shengcheng Dong, with whom I worked on several projects and who taught me the value of patience in scientific research.

I would like to express my appreciation to my committee members for their valuable contributions to my dissertation. The discussions during our committee meetings were engaging and fruitful, and the suggestions and guidance you provided were crucial to the successful completion of my work. I would like to offer a special thanks to Harm Derksen for the stimulating discussions we had about a specific project. From him, I learned a great deal about the importance of rigor in mathematical research.

Furthermore, I would like to express my gratitude to my collaborators at Stanford University for their work on building and maintaining RegulomeDB. I would like to extend special thanks to Ben Hitz, Stuart Miyasato, Emma Spragins, Mingjie Li, Otto Jolanki, and Meenakshi Kagda for their invaluable contributions to the project.

I would like to extend my gratitude to Margit Burmeister and Julia Eussen for their tremendous support during a difficult time in my Ph.D. studies. I cannot thank them enough for their help. I would also like to express my appreciation to Margit Burmeister as my graduate advisor who provided me with great advice on my classwork and career development. Additionally, I would like to thank Julia Eussen for her invaluable support and assistance throughout my Ph.D. studies.

I would like to express my heartfelt gratitude to my dearest friends, Qian Li and Lu Lu. Thank you for always being there for me through this journey. Your presence in my life has been invaluable, and I cannot thank you enough for all that you have done for me.

Finally, I would like to express my deepest gratitude to my family. My parents and grandparents have always valued and supported higher education, even though I am the first Ph.D. in our family. I am incredibly grateful for their unwavering support and encouragement throughout my academic journey. Without their love and support, I could not have achieved this accomplishment alone. They have been my source of strength and inspiration, and I cannot thank them enough for everything they have done for me. Thank you from the bottom of my heart!

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Understanding the functional consequences of regulatory variants is a significant challenge in genomics. Although Genome-Wide Association Studies (GWAS) have provided valuable insights into human phenotypes by identifying genetic variations associated with diseases and complex traits, the functional implications of many of these genetic variants remain unknown, particularly for non-coding regions of the human genome, which account for over 90% of all variants.

To address this challenge, my dissertation focuses on functionally characterizing regulatory elements and their variants in the human genome. Specifically, I define regulatory variants as single nucleotide polymorphisms (SNPs) that can modify the binding affinities of transcription factors (TFs) within the regulatory elements. Such alterations can impact downstream gene expression and potentially contribute to disease progression and trait development. However, characterizing regulatory variants has traditionally relied on the laborious experimental dissection of the human genome, often confined to specific cell types or tissues, thus making it unfeasible to examine all relevant variants in their appropriate biological context. The advent of high-throughput sequencing and computation methods has substantially accelerated the discovery process. In my dissertation, I have developed a series of computational tools and methods to end-to-end characterize regulatory elements and their variants (Fig 6.1).

In Chapter II, I developed a peak calling software, F-Seq2, to accurately define

regulatory element regions from open chromatin assays and ChIP-seq assays. F-Seq2 utilized kernel density estimation and a dynamic “continuous” Poisson test to account for local biases, outperforming state-of-the-art software including MACS2 in terms of precision and recall. Accurate peak calling is essential for downstream analysis, such as differential binding or motif analysis, and lays the foundation for the functional characterization of regulatory variants.

In Chapter III, I advanced a leading regulatory variants database, RegulomeDB, to its second version. RegulomeDB allows users to query variants and obtain a comprehensive list of functional evidence for their variants of interest. The new version of RegulomeDB contains over five times more data than its previous version, providing an even more comprehensive resource for researchers. Additionally, the introduction of a suite of scoring models, namely SURF and TURF, enables accurate summaries of the likelihood that variants function as regulatory variants based on all available evidence.

In Chapter IV, I developed a machine learning model, TLand, as the next version of the RegulomeDB scoring model, to annotate and prioritize regulatory variants in an organ-specific manner. TLand takes advantage of RegulomeDB-derived features and builds a flexible architecture using stacked generalization to reduce overfitting and facilitate future continuous learning. TLand outperformed state-of-the-art models when holding out cell lines or organ allele-specific binding data. By accounting for common data availability issues that often exist in sequence-based deep learning models, TLand accurately prioritized the relevant organs for approximately 2 million GWAS SNPs.

In Chapter V, I introduced a pipeline, Explain-seq, to automatically train and interpret sequence-based deep learning models given genomic coordinates. I demon-

strated the utility of Explain-seq by applying it to a recent STARR-seq dataset to gain insights into enhancer binding patterns in a cell-specific manner. The pipeline identified both known and de novo motifs in the K562 cell line by comparing them to the JASPAR database.

Overall, the computational methods and tools that I developed throughout my dissertation can aid in the discovery and characterization of regulatory elements and variants in the non-coding regions of the human genome.

CHAPTER I

Introduction

Characterizing the functional consequences of variants in the non-coding regions remains a challenge in human genomics. In this dissertation, I developed a series of computational methods to address this challenge. To begin with, I built a peak calling software that accurately maps regulatory regions derived from open chromatin and ChIP-seq assays. I also advanced and extended the leading non-coding regulatory variant database to its second version. I created a machine-learning model capable of predicting regulatory variants genome-wide. Lastly, I created a pipeline to automate the process of training and interpretation of sequence-based deep learning models within the genomics context.

In this chapter, I will first introduce the biological mechanisms that underpin regulatory elements and regulatory variants. I will then describe the high-throughput functional genomic assays and large consortia efforts in making those high-quality datasets accessible. Next, I will discuss the strategies of current regulatory element databases and computational tools to predict regulatory variants. Specifically, I will focus on discussing sequence-based learning models which explore the effects of genetic variants in the human genome. Lastly, I will discuss the limitations of current computational methods to predict regulatory elements and variants.

1.1 Overview of cis-regulatory elements and regulatory variants

My dissertation study focused on the cis-regulatory elements (CREs) located in the non-coding regions of the human genome. These elements, which include enhancers, promoters, silencers, and insulators, have a profound impact on gene expression levels by regulating the binding of transcription factors (TFs) to short DNA sequences known as motifs [1]. The interplay between these TFs and their target regulatory elements is what constitutes a gene regulatory network, which plays a vital role in biological processes such as development, differentiation, and disease progression. However, deciphering these networks presents a challenge since they not only depend on the DNA sequence itself but also on chromatin factors such as accessibility, histone modification, and looping. These epigenetic factors exhibit significant variations between cells, organs, and individuals, making it even more challenging to interpret their functional consequences [1]. In the following sections, I will delve into genome-wide assays that are able to capture chromatin factors activities and map to the human genome to better understand gene regulatory networks.

In addition to characterizing regulatory elements, interpreting the functional consequences of genetic variation within regulatory elements is even more challenging. The fourth chapter of my dissertation focuses on predicting the functional effects of single nucleotide polymorphisms (i.e. SNPs, the most common type of genetic variation) in regulatory elements. SNPs that can alter the binding affinity of TFs to their target sites, subsequently affecting gene expression regulation downstream are referred to as regulatory variants or regulatory SNPs (Fig 1.1). There are statistical association methods developed aiming to pinpoint regulatory variants genome-wide.

Genome-wide association studies (GWAS) identified millions of genetic variants

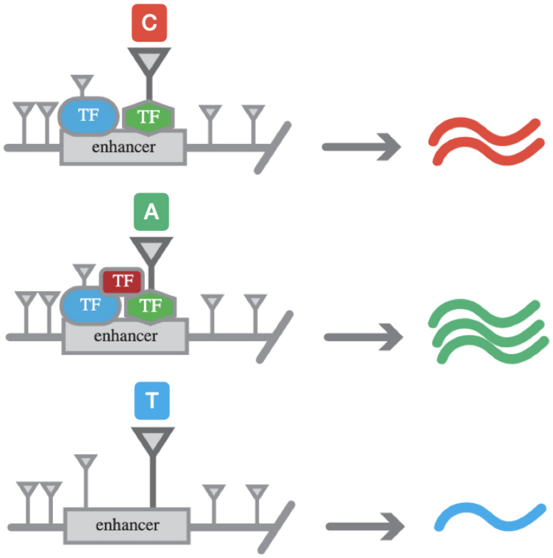


Figure 1.1: Schematic of possible mechanisms for a regulatory single nucleotide polymorphism (SNP) affecting downstream gene expression. The C allele binds with an activator TF to enhance gene expression. A genotype change from C to A may recruit more TFs to further enhance gene expression indicated by the number of wavy lines. However, changing from C to T may disrupt any binding of TFs, thus reducing gene expression. This figure is a revised version of the one originally presented in [2].

that were associated with human diseases and traits [3]. It involves comparing the genomes of individuals with and without the trait or disease of interest to identify genetic differences that may be linked to the trait or disease. For example, to discover variants associated with human height, genetics and height data for individuals were collected. A single variant association test was performed for each candidate variant genome-wide. A multi-test correction was typically conducted to account for a large number of statistical testing being performed. These findings have led to the identification of numerous novel risk loci and a better understanding of the underlying biological mechanisms, with implications for precision medicine. The GWAS Catalog data portal contains over 70,000 variant-trait associations [4]. However, many associations between variants and traits remain unexplained, particularly for the ∼90% of variants in non-coding regions [3]. Linkage disequilibrium among variants is a major cause, making it difficult to pinpoint causal variants for complex traits involving many genes. GWAS loci typically have small effect sizes, and identifying rare causal variants is even more challenging. It is important to note that GWAS only establishes associations between genetic variants and traits or diseases and does not confirm causality. To confirm causal variants and their functions, additional analyses, such as with functional genomic assays, are required.

1.2 High-throughput functional genomic assays to characterize regulatory elements genome-wide

Next-generation sequencing has revolutionized genomics research, providing a range of functional genomics assays to comprehensively study regulatory elements across the entire genome. These assays allow for a detailed characterization of regulatory elements from multiple perspectives (Fig 1.2).

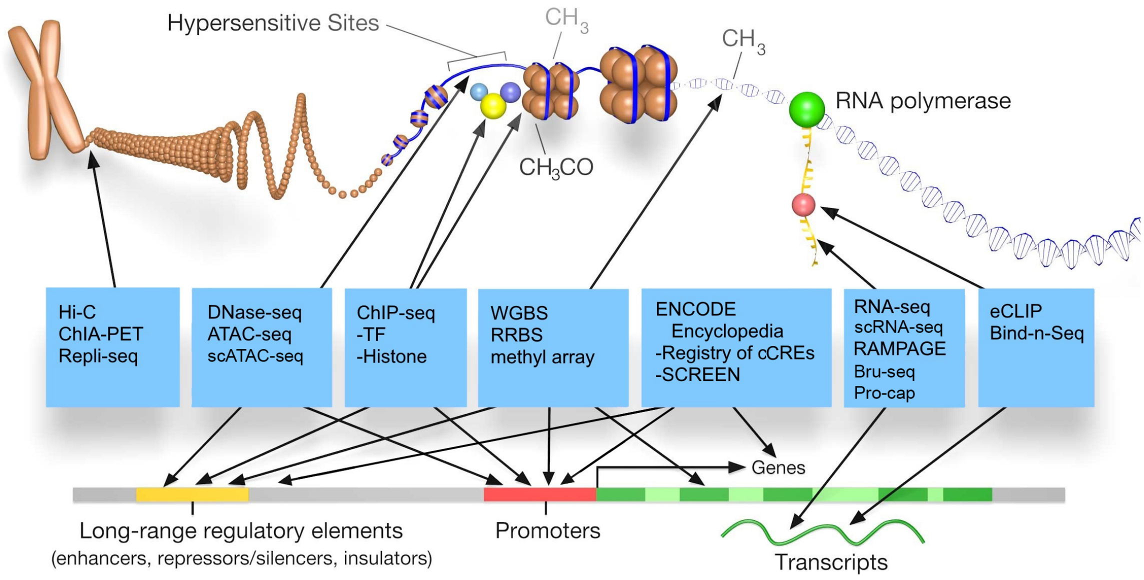


Figure 1.2: High-throughput functional genomics assays characterizing genome-wide regulatory elements from various perspectives. The figure was modified from [5].

1.2.1 Open chromatin assays and footprints

DNase I hypersensitive sites (DHSs) are genomic regions that exhibit increased sensitivity to cleavage by DNase I endonucleases [6]. This characteristic is indicative of a more open and accessible chromatin structure that is accessible to TFs and plays a crucial role in gene regulation. DNase-seq is a sequencing-based technique that leverages DNase I cleavage to identify DHSs across the entire genome [7]. DNase-seq has emerged as a powerful tool for mapping open chromatin regions in different cell types and under varying treatment conditions. ATAC-seq is a more recently developed assay that provides an alternative approach to mapping open chromatin regions [8]. It uses the hyperactive Tn5 transposase to insert sequencing adapters into accessible chromatin regions. While offering similar sensitivity and specificity to DNase-seq, ATAC-seq requires fewer starting cell numbers and preparation steps [8].

Footprints are short regions within open chromatin regions that are bounded by TFs, thus they are protected from the digestion of DNase I or Tn5 transposase. By analyzing the genome-wide signal from open chromatin assays, computational methods are able to identify footprints as the dips within peaks of open chromatin signals, indicating the exact binding site of TFs and implying the underlying sequence as regulatory elements. Some of these methods use sequence information from TF motifs to assign the binding TFs for each footprint [9, 10, 11, 12], while others only map a union set of TF binding sites [13, 14, 15].

1.2.2 Chromatin Immunoprecipitation sequencing (ChIP-seq)

Chromatin Immunoprecipitation sequencing (ChIP-seq) is a method for identifying the binding sites of DNA-associated proteins, such as transcription factors and histone marks, by using antibodies that specifically recognize and pull down the protein of interest from a sample of cells or tissues [16]. The ChIP process involves cross-linking the DNA and protein within cells, breaking the chromatin into small fragments, and then using an antibody specific to the protein of interest to isolate the protein-DNA complexes. The DNA fragments within these complexes are then sequenced using next-generation sequencing technologies, allowing for the identification and mapping of the genomic regions that are bound by the protein of interest [16]. By comparing ChIP-seq data from various different conditions, we could gain insights into the mechanism of gene regulatory networks and other biological processes.

Before mapping the sequence data obtained from ChIP-seq experiments to a reference genome, a quality check is usually performed to filter out reads with low-quality scores. Peak calling software is then employed to identify regions of the genome that are enriched for the protein of interest. The accuracy of different peak calling software algorithms is crucial for downstream analysis, such as differentially binding analysis and motif enrichment analysis [17]. Therefore, choosing a peak calling method that can provide reliable and accurate results is important to ensure the validity of subsequent analyses.

1.2.3 Peak calling software to identify regulatory regions

Peak calling algorithms work by analyzing the read coverage of the sequencing data obtained from the genomic assay. Regions of the genome with a high density of reads are identified as peaks and are presumed to represent the locations of the protein-DNA interactions of interest. Many peak calling programs have been developed, each with its own strengths and weaknesses (see Table 1.1), mostly due to their various approaches to handling/modeling sequencing reads [18, 19, 20, 21, 22, 23, 24, 17, 25, 26, 27, 28, 29, 30, 31].

Table 1.1: Comparison of common peak calling software. Columns are common peak calling software. Rows are various characteristics that can be grouped into two categories, locating the potential peaks and ranking of peaks. Green indicates certain peak callers have specific characteristics. The table was modified from [32].

	GEM	BCP(TF)	BCP(Hist)	MUSIC	MACS2	SICER	SISSRs	F-Seq	F-Seq2	Hotspot	spp-wtd	spp-mtc	spp-msp	CisGenome	TM	QuEST	PolyPeak	Qeseq
Locating the potential peaks																		
High Resolution	🟩	🟩		🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩	🟩
Chip and input sample signals combined						🟩									🟩			🟩
Multiple alternate window sizes	🟩	🟩	🟩	🟩					🟩									
Use of variability of local signal	🟩	🟩	🟩		🟩	🟩		🟩	🟩	🟩				🟩		🟩	🟩	🟩
Ranking of peaks																		
Binomial test	🟩			🟩		🟩				🟩				🟩				
Poisson test		🟩			🟩	🟩	🟩		🟩		🟩	🟩	🟩					
Normalized difference score															🟩			
Use of underlying genome sequence	🟩																	
Fold-change																🟩		🟩
Comparion of distributions																		🟩
Posterior probability of enrichment			🟩															
Shape based statistic																	🟩	

The accuracy and reliability of peak calling software are crucial for downstream analyses. Therefore, it is important to carefully evaluate the performance of different peak calling methods to ensure that the results obtained are both valid and biologically meaningful. Note that they may perform differently depending on the specific dataset and research question. In the second chapter of my thesis, I will introduce our most recently developed peak caller, F-Seq2, and demonstrate its superior performance over the state-of-the-art software on various genomic assays with corresponding software settings.

1.3 Functional genomic data resources and databases

1.3.1 Large genomic consortia

Creating a comprehensive human regulatory map through individual efforts is limited. Additionally, individuals may adopt different approaches to analyzing data, making it difficult to perform comparative analyses and/or draw conclusions from multiple data sources. Large consortia have been established as collaborative efforts among multiple research institutions and organizations, such as the Encyclopedia of DNA elements (ENOCDE) consortium [5], the Roadmap Epigenomics Mapping Consortium [33], and the IGVF (Impact of Genomic Variation on Function) Consortium [34] which is the successor of ENCODE. These consortia bring together diverse expertise, resources, and data to accelerate scientific discoveries, identify disease risk factors, develop new treatments, and advance precision medicine. Uniform processing pipelines and integrative analysis methods are often developed to ensure standardized, comparative, and non-redundant results.

As of March 2023, ENCODE hosts 14,805 assays, including over 3,000 TF ChIP-seq assays, over 3,000 histone ChIP-seq assays, and approximately 1,500 DNase-seq assays across cell lines and conditions [5]. ENCODE's focus is on human cells grown in culture and primary human tissues and cells, while Roadmap focuses on samples taken directly from human tissues and cells. In contrast, Roadmap hosts a comprehensive collection of 1,320 datasets for 127 consolidated epigenomes, consisting of 105 DNA methylation datasets and 127 consolidated ChIP-seq datasets for a core set of histone modifications (i.e. H3K4me1, H3K4me3, H3K27me3, H3K36me3, and H3K9me3) [33].

1.3.2 Cis-regulatory elements databases

Large consortia work in genomics has greatly advanced our understanding of the genetic basis of complex diseases. However, these efforts often focus on cataloging large datasets of genomic assays. Efficiently querying and annotating specific regulatory elements and variants with comprehensive datasets is needed (Fig 1.3). Cis-regulatory element databases and web portables have been developed to annotate regulatory elements and prioritize regulatory variants, for example, SCREEN (Search Candidate cis-Regulatory Elements by ENCODE) [5], FAVOR (Functional Annotation of Variants) [35], and RegulomeDB [36] which I will mainly discuss in the third chapter.

Integrative analysis is a key aspect of such databases where they integrate multiple ground-level annotations (e.g. peaks and quantifications for individual data types produced by the ENCODE uniform processing pipelines) and generate new annotations or summarization scores. SCREEN implemented a Z-score to define high epigenomic signal regions as candidate regulatory elements. Specifically, SCREEN started with 93 million individual DHSs across 706 DNase-seq assays in human datasets. For each DHS, the Z-scores were computed as the log10 of DNase, H3K4me3, H3K27ac, and CTCF signals in each biosample data. In total, SCREEN's current version 3 identified 926,535 human candidate cis-regulatory elements which have high DNase Z-scores and high H3K4me3, H3K27ac, and/or CTCF ChIP-seq Z-scores (high score

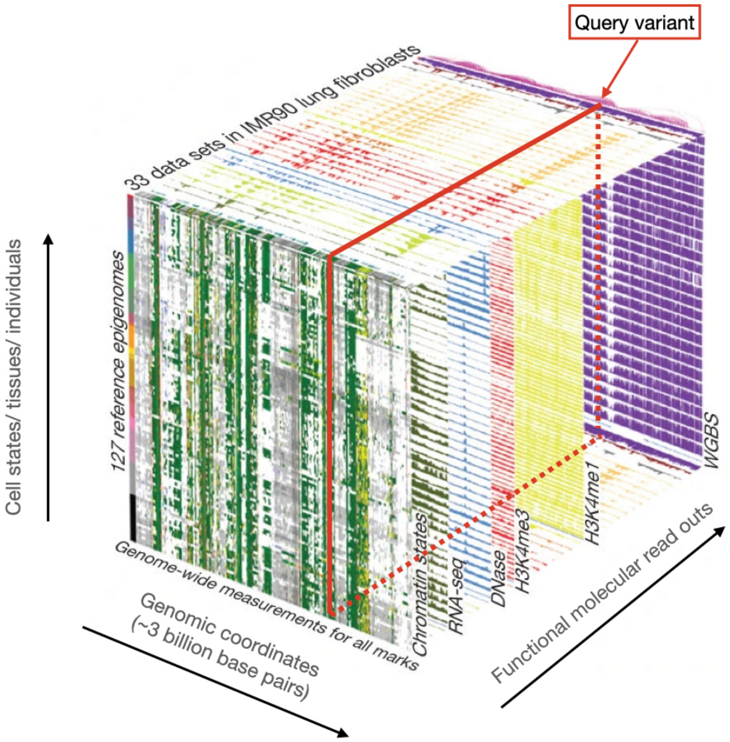

Query variant
33 data sets in IMR90 lung fibroblasts
127 reference epigenomes
Cell states/ tissues/ individuals
Genome-wide measurements for all marks
Genomic coordinates
(~3 billion base pairs)
Chromatin states
RNA-seq
DNase
H3K4me3
H3K4me1
WGBS
Functional molecular read outs

Figure 1.3: Query regulatory variant information from large consortia databases. The figure was modified from [5].

defined as Z-score > 1.64, corresponding to the 95th percentile of a one-tailed test).

In contrast, FAVOR employed a single metric summarizing multiple similar annotations measuring the same underlying biological function (variant annotation Principal Components, aPCs). They are the principal components summarizing the multifaceted functional annotation data in FAVOR. For example, aPC-Epigenetics-Active was defined as the first PC of the standardized scores of EncodeH3K4me1.max, EncodeH3K4me2.max, EncodeH3K4me3.max, EncodeH3K9ac.max, EncodeH3K27ac.max, EncodeH4K20me1.max, EncodeH2AFZ.max, in PHRED scale [37]. FAVOR version 2 contains annotations and scores for a total of 8,892,915,237 variants (including 8,812,917,339 SNVs and 79,997,898 indels).

On the other hand, RegulomeDB provides a genome-wide regulatory variant score map based on a suit of machine learning models, SURF and TURF [38, 39]. The regulatory variant score incorporates all the datasets information from ENCODE, Roadmap, and the Genomics of Gene Regulation Consortium [40], representing the likelihood of a variant having any regulatory function. RegulomeDB web portal allows users to query any variants genome-wide in either the GRCh38 or hg19 genome assembly via rsID or genome coordinates.

Active maintenance and development is another key aspect of such databases. This allows databases to leverage the continuously growing wealth of genomic data and provide a valuable resource for researchers to investigate the functional impact of genetic variation and identify novel therapeutic targets for complex diseases. Keeping up with high-quality datasets is essential for researchers to screen for new targets. To achieve this, RegulomeDB mirrored the ENCODE portal to stay current with the latest data. Recently, RegulomeDB has been advanced to its second version, including five times more data than the previous version. The database is being actively maintained and developed, with new scoring models TLand models, being developed to prioritize regulatory variants in an organ-specific manner. TLand models would be available in the upcoming third version.

1.4 Computational methods to predict regulatory elements and variants

Computational methods have been developed for predicting regulatory elements and variants in genomics. As a data-driven field, genomics largely utilizes machine learning, such as random forest models in SURF and TURF, to uncover dependencies in data and generate innovative biological hypotheses. In this section, I will first focus on sequence-based learning as an important approach to studying regulatory regions. I will explore its advantages and disadvantages, as well as potential avenues for improvement. Finally, I will examine the general limitations of current computational methods in this field.

1.4.1 Sequence-based machine learning models

The ability to predict genomic signatures (e.g. gene expression, chromatin states, and TF binding sites) purely from DNA sequences have significant potential to advance our understanding of gene regulatory networks and their impact on human diseases and traits. Noncoding genetic variants associated with human diseases and traits can be difficult to understand and study through population-based association studies such as GWAS, which are often limited to common variants and struggle to disentangle causality from association due to linkage disequilibrium (LD). Moreover, the experimental validation of human genetic variants is a time-consuming and difficult process, limited to cell types or tissues that can be accurately replicated in the laboratory. As a result, testing all relevant variants of interest in the relevant biological contexts is often impractical.

Support vector machine

Sequence-based machine learning models have the potential to overcome these limitations and allow for a more comprehensive understanding of the complex regulatory mechanisms underlying human diseases and traits. In particular, Support Vector Machine (SVM) models have proven to be a useful tool in genomics due to their ability to handle high-dimensional data and non-linear relationships between variables. One type of SVM model, gkm-SVM, has been specifically designed for predicting TF binding sites [41]. Authors developed a novel gapped k-mer (i.e. short sequences of DNA) based approach to train models. The advantages of gkm-SVM models include their ability to handle variable length input DNA sequences and capture non-linear dependencies between k-mers in an arbitrarily high dimension.

Gkm-SVM significantly outperformed other state-of-the-art methods for predicting regulatory elements as TF binding sites [41].

Convolutional neural networks

However, to effectively extract new insights from the rapidly growing volume of genomics data, more expressive machine learning models are required. While traditional machine learning methods like SVM have proven accurate in medium datasets (e.g. 10,000 samples), the advent of deep learning has transformed fields like computer vision and natural language processing by effectively leveraging large datasets. Deep learning is becoming the preferred method for many modeling tasks, including predicting the impact of genetic variation on gene regulatory mechanisms such as DNA accessibility and splicing.

Convolutional neural networks (CNNs) have proven particularly useful in genomic sequence learning tasks because their convolutional filters resemble position weight matrices (PWMs). For instance, when identifying if genomic regions are bound by a specific transcription factor (TF) indicated by its ChIP-seq peak regions, k-mers or PWMs representing the TF binding site patterns are often employed to scan sequences for matches. This is because TFs bind to DNA by recognizing sequence motifs or patterns that are resistant to shifts within the sequences. However, patterns in which transcription factor binding depends on a combination of multiple motifs with well-defined spacing would not be accurately recognized by PWMs or k-mers [42]. In addition, the number of potential k-mers increases exponentially with k-mer length, leading to storage and overfitting issues. On the other hand, a convolutional layer in a CNN enforces translational invariance by applying the same convolutional filters to every position in its input sequence. This process can be thought of as scanning the sequence using multiple PWMs, thereby allowing for the more effective capture of complex dependencies among TFs [42, 43].

DeepSEA [44], Basset [45], and DeepBind [46] were the pioneer CNN models developed within the genomics context. The DeepSEA model takes input as a 1,000 bp DNA sequence to predict the presence of 919 (epi-)genomic features, including open chromatin, TF binding sites, and histone modifications. Its successor, the Sei model, simultaneously predicts 21,097 features with a larger receptive field of 4,000 bp sequence [47]. Basset predicted 164 binary targets of open chromatin features given a 600 bp input sequence. All CNN models substantially outperformed the gkm-SVM models.

Recurrent neural networks and transformers

Regulatory elements can regulate genes distally by forming 3D loops. Recurrent neural networks (RNNs) are able to carry over information through infinitely long sequences via memory theoretically, which is very suitable to be employed to model such long-range dependencies. In addition, RNNs can take input of widely varying lengths rather than taking uniform length input as CNNs. RNNs have been shown to perform better than CNNs with the same training and test datasets in predicting open chromatin, TF binding sites, and histone modifications [48, 49]. However, recent systematic comparisons have demonstrated that CNNs can achieve comparable or even superior performance compared to RNNs in sequence modeling tasks, such as audio synthesis and machine translation, when combined with various techniques, including dilated convolutions [50]. Furthermore, RNNs apply a sequential operation that hinders their parallelization, resulting in much slower computation compared to CNNs.

A new deep learning architecture, transformer, has made substantial breakthroughs in the natural language processing field [51]. Transformers are built with attention layers that transform each position in the input sequence by computing a weighted sum across the representations of all other positions in the sequence. This is achieved by considering the embeddings of their current representation vectors and the distance between them, which enable models to learn long-range dependencies efficiently. For example, to predict a gene expression level, transformers can collect information on distal regulatory elements away from the gene, while CNN requires multiple successive layers to reach the same distal regions due to its local receptive field. Recently, transformer-based models have been developed in genomic sequence learning tasks, such as Enformer [52] and DNABERT [53]. Notably, Enformer is able to model distal regulatory elements up to 100 kb away, while the previous state-of-the-art CNN models can only reach up to 20 kb.

Despite the effectiveness of large language models such as Enformer in modeling the distal dependencies, there are still relationships to be captured which can not be solely solved by increasing the range of sequence learning. Trans-regulatory elements can regulate genes on different chromosomes as they encode proteins or other molecules that can diffuse through the cell and interact with DNA sequences on different chromosomes. Sequence learning has its upper ceiling limit in predicting genomic signatures since these cross-chromosomal relationships, for example, are challenging to model. Modeling with multiple modalities rather than solely relying on genomic sequences gains a more comprehensive understanding of the gene regulatory networks.

1.4.2 Integrating multimodal methods

Data integration methods can be broadly categorized into three types for developing a final comprehensive method to study regulatory elements and variants (Fig 1.4) [54]. Early integration (often referred to as data concatenation) involves transforming all datasets into a single feature-based representation, which is used as input to a final method such as a machine learning model. Projection methods, such as dimensionality reduction, are often used in early integration to concatenate low-dimensional representations of high-dimensional data. Late integration (often referred to as model ensemble), on the other hand, involves building base-level models for each dataset or data type independently and then combining their predictions using methods such as majority voting and averaging with weights. Alternatively, one can train a meta-model to learn the best weights for combining base-level models using stacked generalization [55]. Multiple layers of base-level models can also be built to capture more complex dependencies.

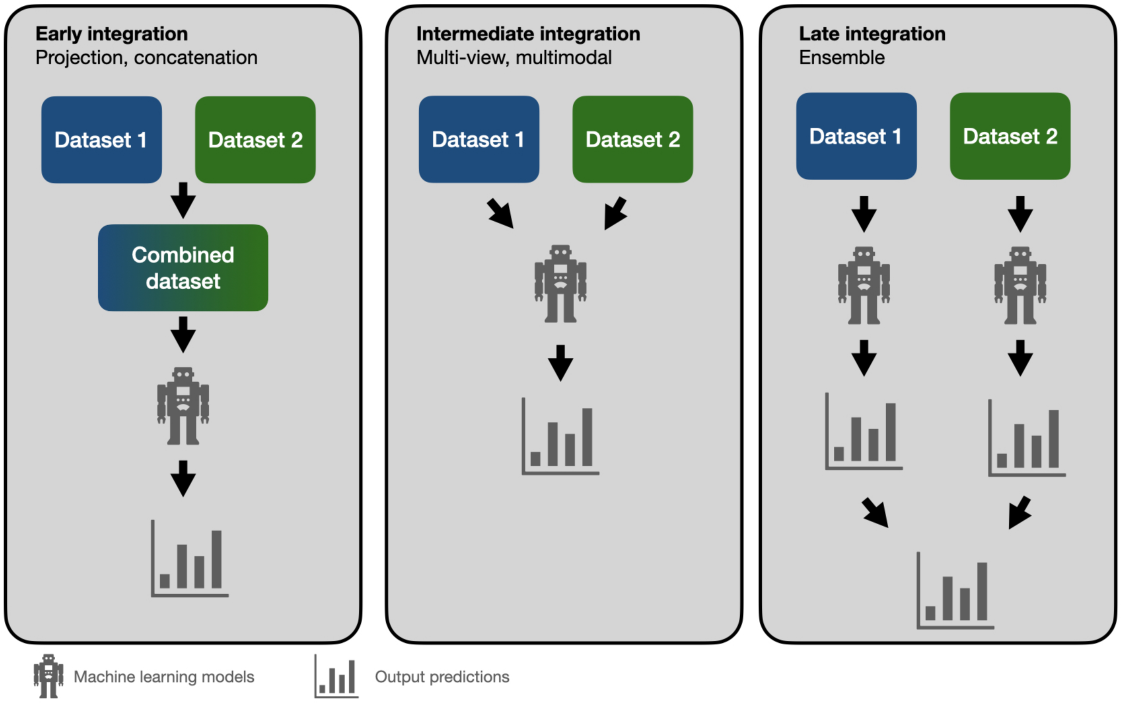


Figure 1.4: Categorization of data integration methods. Data integration approaches can be divided into three categories, early integration, intermediate integration, and late integration. The figure was modified from [54].

Intermediate integration, also known as multi-modal learning, is another type of data integration method. For example, to develop a final deep learning model, intermediate integration involves learning a joint representation of multiple datasets by developing different layers that explicitly address the multiplicity of datasets and fuse them through inference of a joint layer. Intermediate integration preserves the structure of data and only merges during the analysis stage, leading to superior performance over the other two data integration methods [54].

1.4.3 Challenges for current computational methods

Advancements in machine learning and emerging applications are opening up exciting possibilities for understanding human non-coding regulatory elements. However, recent studies have shown that there is no one-size-fits-all approach when it comes to selecting the best method for a given problem [54, 56]. To achieve optimal results, methods must be selected based on specific data types, domain models, and research questions.

Recognizing that a single method may not suffice has led to a focus on integrating multiple methods to increase accuracy or scale. For example, by combining a comprehensive set of cell-specific models, a multi-scale model for an organ, or even for an organism, may capture the full extent of biological complexity. Nevertheless, integrating approaches are still in the early stages, and the key principles of optimal design are not yet fully understood.

Another challenge is the limited generalizability of current methods to new conditions [52]. Models may perform well in the context where they are trained, such as a specific cell line or transcription factor, but their transferability to other conditions remains a cutting-edge research topic. To enable transferability, models must be trained on a comprehensive set of input features and learn high-level patterns that can be applied to other conditions, such as another cell line in humans. This also requires easy access to a comprehensive database for generating the necessary features.

In the third and fourth chapters of my dissertation, I will introduce our approach to navigating those challenges by developing a comprehensive yet easily accessible database and developing an ensemble machine learning model to prioritize human regulatory variants in an organ-specific manner.

CHAPTER II

Calling Peaks to Define Regulatory Regions in the Human Genome

2.1 Abstract

Genomic and epigenomic features are captured at a genome-wide level by using high-throughput sequencing (HTS) technologies. Peak calling delineates features identified in HTS experiments, such as open chromatin regions and transcription factor binding sites, by comparing the observed read distributions to a random expectation. Since its introduction, F-Seq has been widely used and shown to be the most sensitive and accurate peak caller for DNase I hypersensitive site (DNase-seq) data. However, the first release (F-Seq1) has two key limitations: lack of support for user-input control datasets, and poor test statistic reporting. These constrain its ability to capture systematic and experimental biases inherent to the background distributions in peak prediction, and to subsequently rank predicted peaks by confidence. To address these limitations, we present F-Seq2, which combines kernel density estimation and a dynamic "continuous" Poisson test to account for local biases and accurately rank candidate peaks. The output of F-Seq2 is suitable for irreproducible discovery rate (IDR) analysis as test statistics are calculated for individual candidate summits, allowing direct comparison of predictions across replicates. These improvements significantly boost the performance of F-Seq2 for ATAC-seq and ChIP-seq datasets,

outperforming competing peak callers used by the ENCODE Consortium in terms of precision and recall.

2.2 Introduction

High-throughput sequencing (HTS) is a central technology in deciphering genomic and epigenomic landscapes. Assays for detecting genome-wide chromatin accessibility [7, 8, 57], transcription factor (TF) binding [16], and histone modifications [58] are among the most commonly used methods. The short read sequences produced by these assays are usually filtered and mapped back to a reference genome, then accumulated and piled up in genomic regions. The enrichment (e.g. counts) of mapped reads can be abstractly viewed as a digital signal of relevant biological events varying along the genome. The genome-wide enrichment signal can be further processed with a peak-calling program, or peak caller, to find the arguments of local maxima (argmax), representing discrete loci with statistically significant enrichment over background for the relevant biological event. For example, individual TF binding sites in a ChIP-seq experiment.

We introduced F-Seq as a general peak caller for DNase-seq and ChIP-seq in 2008 [24]. Unlike other recent methods [21, 20], F-seq calls peaks in HTS signals which are the probabilistic estimates of the genome-wide short read density at single-nucleotide resolution reconstructed by a kernel density estimator (KDE) [59, 60]. KDE-based reconstructed signal is smoother and more accurate than histogram-based methods (e.g. sliding window), but still interpretable and useful for visualization as the estimate is proportional to the probability of finding a read at a given base pair [61]. A Gaussian kernel with a chosen bandwidth is centered at each read and kernels are summed up to obtain the density estimate. Peak regions are then called if the signal is higher than the threshold calculated from a simulated background model. F-Seq has been widely used in the ENCODE project [5] and beyond, which is shown to be more accurate and sensitive than competing peak callers for DNase-seq data [62]. However, F-Seq lacks native support for a separate control dataset. Consequently, F-Seq cannot capture or eliminate local biases affecting read distribution along the genome, such as copy number variation, read mappability, and local chromatin structure [21]. This limits the performance of F-Seq especially on ChIP-seq data since the majority of ChIP-seq experiments have corresponding control data which contains unique information for accurate peak calling [63]. In addition, F-Seq does not report test statistics (e.g. p-value or q-value) apart from the signal value at each position.

To address these shortcomings, we have developed F-Seq version 2 (F-Seq2), a complete rewrite of the original F-Seq in Python. F-Seq2 implements a dynamic parameter to conduct local statistical analysis with an underlying "continuous" Poisson distribution which is approximated by logarithmic interpolation of p-values. This allows a Poisson test for continuous signal values (i.e. amplitude) at each genomic position to the local background distribution. By combining the power of the local test and the KDE, which model the read probability distribution with statistical rigor, we robustly account for local biases and solve ties that occur when ranking candidate summits, making results suitable for irreproducible discovery rate (IDR) analysis [64]. We compared F-Seq2 with four peak callers used by the ENCODE Consortium [5] on simulated and real ChIP-seq and ATAC-seq datasets, demonstrating performance gains arising from the joint effect of KDE and the local test, especially in the absence of control data.

2.3 Methods

2.3.1 Density profiles and peak calling

Density profiles for HTS reads at any base pair position x of the genome are defined as

$$\hat{\rho}(x) = \frac{C}{b} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right)$$

where $K(x) = \frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}}$ is a Gaussian kernel density function, and b is the bandwidth parameter controlling the smoothness of the estimation. In contrast to the original KDE, F-Seq2 density profiles represent unnormalized estimates (i.e. not normalized to the total read count) for computational convenience of following statistical analysis. C is a scaling constant so that the sum at any given position is limited to the number of proximal sample points. For experiments including a control dataset, scaling between control and treatment datasets was necessary to account for different sequencing depths. The total control read count was linearly scaled to be equal to the total treatment read count at the individual chromosome level as the ratios of total reads fluctuated between different chromosomes. The reconstructed signal by KDE was treated as a digital signal emitted on a chromosome. Argmax of the signal, which are the positions of local maxima in the estimated density function, were established by comparing neighboring values. Only a subset of argmax were retained as the candidate summits for statistical testing to reduce potential false positives. Candidates were selected by their local maxima properties; we specified the minimum height and prominence of the local maxima for candidates as the simulated background threshold and the minimum distances between adjacent local maxima as the estimated fragment size. Estimation of the fragment size for ChIP-seq data, and the simulated background threshold for defining and selecting candidate summits and delineating final peak regions were implemented the same between F-Seq2 in Python and the original F-Seq in Java.

We adopted and modified the dynamic testing idea introduced by MACS2 [21] to assign each candidate summit a statistical enrichment value related to a background distribution. Rather than using a constant background estimation for all candidates, a local background distribution was estimated for each candidate, providing a more accurate method to calculate enrichment p-values due to the local fluctuations of read enrichment distributions. The Poisson distribution (characterized by) was used to model the number of reads (or signal value) from a genomic region as this has been proven to be more mathematically powerful compared to Binomial distribution in peak calling [32]. Specifically, λ for a summit is defined as $\lambda_{\text{local}} = \max\left(\lambda_{BG}, \left[\lambda_{p1}, \lambda_{1k}\right], \lambda_{5k}, \lambda_{10k}\right)$, where λ_{p1} is the maximum signal value for one pseudo-read, λ_{BG} is the estimate of the individual chromosome background, and λ_x is the estimate of a x bp window centered at the summit. All estimates are calculated in the control dataset where available; otherwise, estimates were only calculated in the treatment dataset, and regions in the square brackets of formula were excluded to alleviate the background estimation boost by the summit signal value.

Since the underlying Poisson distribution of the statistical test is a discrete distribution while the test sample (i.e. the signal value) is continuous, many ties in test statistics p-value calculated by survival function were observed. Supposing $X \sim \text{Pois}(\lambda)$, the Poisson survival function is then defined as $S(X = x; \lambda) = 1 - \sum_{i=0}^{x} \frac{\lambda^i e^{-\lambda}}{i!}$. Ties often occurred when the sequencing data had a low signal-to-noise ratio and KDE estimated signal values were close to each other (i.e. between two integers), such as $S(2.1, \lambda) = S(2.9, \lambda) = S(2, \lambda)$. We interpolated the p-value in the logarithmic space of the survival function to allow for continuous input, and break any ties that occurred. The interpolated p-value in logarithmic space is calculated as

$$\log_{10}(\hat{S}(Y = y; \lambda)) = (y - \lfloor y \rfloor) \cdot \log_{10}\left(\frac{S(\lceil y \rceil; \lambda)}{S(\lfloor y \rfloor; \lambda)}\right) + \log_{10}(S(\lfloor y \rfloor; \lambda))$$

where Y is a continuous random variable, $\lfloor y \rfloor$ is the floor function, and $\lceil y \rceil$ is the ceiling function. The precision gained by this interpolation improved the rankings of summits compared to the rankings calculated using discrete values. The interpolation bridges KDE and the dynamic Poisson testing to combine their power. Multi-test correction was conducted with the Benjamini-Hochberg approach [65] to calculate q-values (more precisely, false discovery rate adjusted p-values) from the interpolated p-values.

2.3.2 Benchmarking with selected peak callers

Four peak callers and F-Seq2 were selected to benchmark our improved method on 100 simulated HTS datasets, 3 real ChIP-seq datasets, and one ATAC-seq dataset. The comparison methods, which are routinely utilized by the ENCODE Consortium [5], included Model-based Analysis for ChIP-Seq version 2 (MACS2) [21], SPP [26], MUltiScale enrIchment Calling for ChIP-Seq (MUSIC) [20], and Genome wide Event finding and Motif discovery (GEM) [18]. 100 treatment datasets and their paired control samples were simulated to closely approximate real ChIP-seq datasets [32], allowing for the evaluation of the peak callers under different scenarios where the ground truth is known. Real ChIP-seq datasets for 3 different TFs tested in 3 different cell lines were obtained from ENCODE [5]. As the ground truth is unknown in real datasets, one common approach is to use the presence of a matched TF binding motif to indicate true positive peak predictions. Motifs were obtained from the JASPAR database [66] irrespective of cell line specificity, and used for the 3 real ChIP-seq datasets. Similarly, the union set of conservative IDR peaks from 117 independent ENCODE TF ChIP-seq experiments were used as the “ground truth” for ATAC-seq benchmarking [5]. Raw ATAC-seq bam files were downloaded from Buenrostro et al. [8] (see Availability for data accession numbers).

Performance for all peak callers was evaluated across a range of significance thresholds representing a different number of top ranked peaks. The main evaluation metric was the F-score defined as

$$F_\beta = \left(1+\beta^2\right) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

$$\text{precision} = \frac{tp}{tp+fp}$$

$$\text{recall} = \frac{tp}{tp+fn}$$

When $\beta = 1$, we refer to it as F-score, or more specifically, F1-score; when $\beta = 0.5$, we refer to it as F0.5-score. tp is the number of true positives, fp is the number of false positives, and fn is the number of false negatives. A higher F-score indicates a more balanced performance in terms of precision and recall. All peak callers were run with recommended settings and the least stringent thresholds (i.e. set p-value or q-value threshold to 1 or fold enrichment threshold to 0; see Availability for parameters settings).

2.3.3 Evaluation for simulated data

Peak calling results are typically not directly comparable as they possess different peak widths and estimated p-values or q-values that are generated from different statistical tests. To address this issue, all tools were first run with the least stringent threshold to obtain an extensive list of peaks on each simulated dataset for each tool. All peaks were limited to a 200 bp window centered at the peak summit or peak centers, depending on available dataset information. Operating characteristics can be evaluated by varying the threshold to obtain the same top number of peaks from each tool, where peaks are ranked by individual significance measurements. F-score was used as the evaluation metric, which is the harmonic mean of precision and recall. Specifically in the simulation evaluation, tp was defined as the number of predicted peaks which overlap with ground truth peaks. $Precision_{simulation}$ was defined as the fraction of tp in all predictions, and $recall_{simulation}$ as the fraction of tp in all ground truth peaks. The mean and 95% confidence intervals across 100 peak calling results were estimated by generalized additive models (GAMs) [67] for each peak caller. A linear GAM was fit to the results for regression analysis. Using the fitted model to predict on the varying threshold generated the mean curve and 95% prediction intervals, which was defined as 95% confidence intervals for each peak caller. Since the lengths of the operating characteristics curves varied due to the different maximum number of peaks called by each peak caller, and different p-values or q-values sensitivities responding to the varying threshold, the area under the curve statistics used to summarize the curve were not directly comparable. We then used the highest F-score and the overall trend of the curve for peak caller evaluation. The higher the overall curve, the larger the area under the curve, the more balanced and optimal the performance of a peak caller is in terms of precision and recall.

2.3.4 Evaluation for ATAC-seq data

Evaluation of F-Seq2 and MACS2 used the union set of conservative IDR peaks from 117 TF ChIP-seq datasets as the “ground truth”. All IDR peaks were in the GM12878 cell line to be comparable to the ATAC-seq dataset. Each tool was run with the least stringent threshold and two main modes: single-end (SE) and paired-end (PE) mode. Paired-end mode has the advantage of knowing the exact fragment length, which is useful when filtering out fragments whose length falls within a certain range to avoid peak calls on nucleosome centers [8]. Operating characteristics curves were plotted similarly as described in the evaluation for simulation data by varying the respective thresholds. The main difference was the evaluation metric was changed to F0.5-score along with new definitions for true positives, precision and recall. We used F0.5-score to put more emphasis on precision versus recall due to the incompleteness of the "ground truth". tp was redefined as the number of base pairs (bp) of the predicted peaks that overlap with ground truth peaks, $Precision_{atac}$ as the fraction of correctly predicted base pairs in all predictions, and $recall_{atac}$ as the fraction of correctly predicted base pairs in all ground truth peaks. New definitions were required as ATAC-seq peak lengths are usually larger than TF ChIP-seq peak lengths. We shifted focus from evaluating summits around a window size to the narrow peak regions for a more comprehensive evaluation.

2.3.5 Evaluation for real TF ChIP-seq data

Evaluations of real TF ChIP-seq peak calling results required JASPAR motif Position Weight Matrices (PWM) of each TF. K-mers matching to each TF PWM were identified by the TFM P-value program [68] with the threshold of 4^{-8}. Motif positions were detected in the hg19 human genome by mapping the K-mers using the Bowtie program suite [69]. For each ChIP-seq dataset, the selected tool called a list of significant peaks with their default thresholds. The shortest distances between the significant peaks and the corresponding TF motifs were obtained and used as the main evaluation metric. Specifically, we evaluated the fraction of top n up to 1000 peaks, ranked by significance within a 100 bp window of a motif. We also examined the empirical cumulative distribution of the shortest distance of those top 1000 peaks for each tool.

2.3.6 F-Seq2 auto filter design for paired-end ATAC-seq data peak calling

We designed the PE auto filter based on the fragment size distribution partitions modelled by Buenrostro *et al.* [8], where fragment lengths under 100 bp, between 180 and 247 bp, between 315 and 473 bp, and between 558 and 615 bp were considered to originate from nucleosome free, mono-, di-, and tri-nucleosomes, respectively. Our auto filter included more fragments compared to that of Buenrostro's analysis [8], in which they only used fragments under 100 bp for open chromatin analysis (Fig 2.1). By excluding fragment ranges between the non-overlapping cutoffs, a large percentage (∼15%) were discarded, leading to a reduction in recall. These fragments (e.g. between 100 and 180 bp) may contain useful information for identifying open chromatin regions [70]. F-Seq2 takes advantage of more available reads to accurately estimate background distribution, and only fragments within mono-, di-, and tri-nucleosomes ranges were excluded. Fragments larger than 558 bp (i.e. multinucleosome-sized fragments) were also rejected as these fragments are associated with condensed heterochromatin [8].

2.4 Results

2.4.1 Performance on simulated datasets

To accurately evaluate the peak callers under a variety of scenarios, each method was benchmarked on 100 sets of paired simulated treatment and control data. F-Seq2 and MACS2 were found to be the top two performers with the highest overall F-score operating characteristic curves (Fig 2.2A). The highest F-scores estimated by generalized additive models across 100 pairs were 0.897, and 0.884 for F-Seq2 and MACS2, respectively. Both methods outperformed MUSIC, the third-best method, by a margin of ∼0.1 (MUSIC 0.781). Despite differences in implementing a dynamic

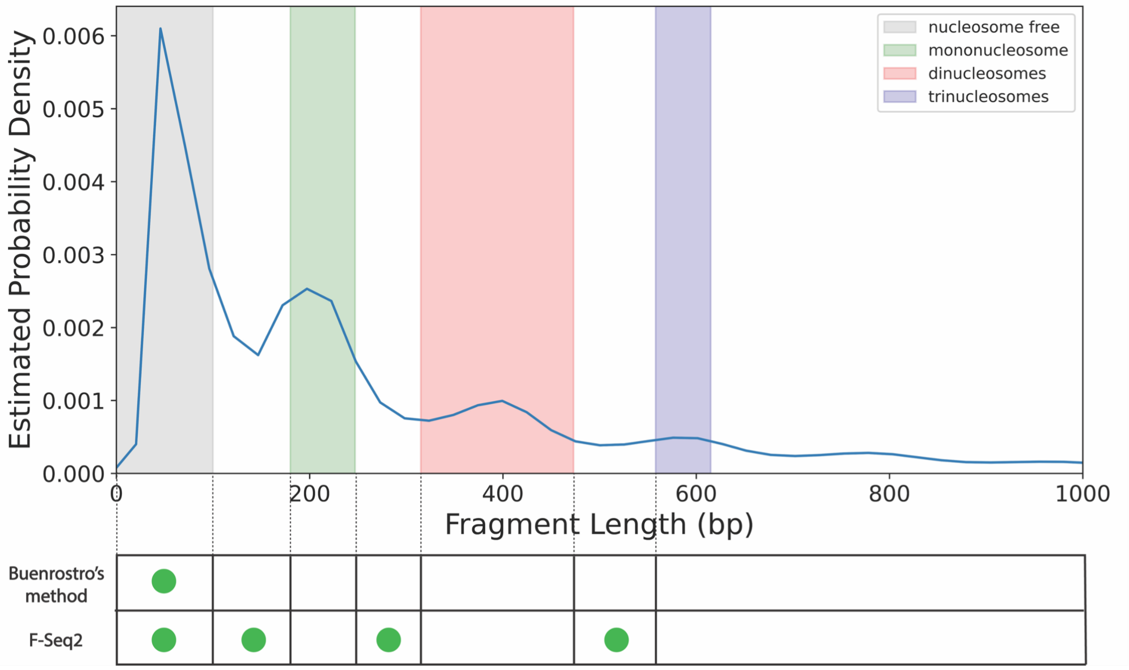


Figure 2.1: Fragment size distribution of the ATAC-seq datasets in GM12878. In Buenrostro's study, the observed fragment distribution was partitioned into four populations of reads, including nucleosome free, mono-, di-, and tri-nucleosomes, by fitting the distribution to one exponential function and five Gaussians (2). The probability density was estimated by KDE. Green dots show the populations of reads included in each method to call peaks on open chromatin regions. Users can adjust the boundaries in the F-Seq2 program for a better fit of four populations after observing their fragment size distribution.

parameter λ_{local} between F-Seq2 and MACS2, the performance gap suggests using a dynamic parameter λ_{local} in ranking peaks is a huge advantage, effectively removing false positives, consistent with the conclusion from Thomas *et al.* [32]. The number of peaks called by the default threshold of each peak caller was compared to the number of peaks in the ground truth (Fig 2.2B). F-Seq2 best correlated with the ground truth (r=0.88) while MUSIC (r=0.74) had a slightly better correlation compared to MACS2 (r=0.70). The high correlation observed for F-Seq2 indicates the default threshold of our program is reliable when estimating the number of significant peaks under a simulation setting.

Although control data is often essential for modeling background distributions for

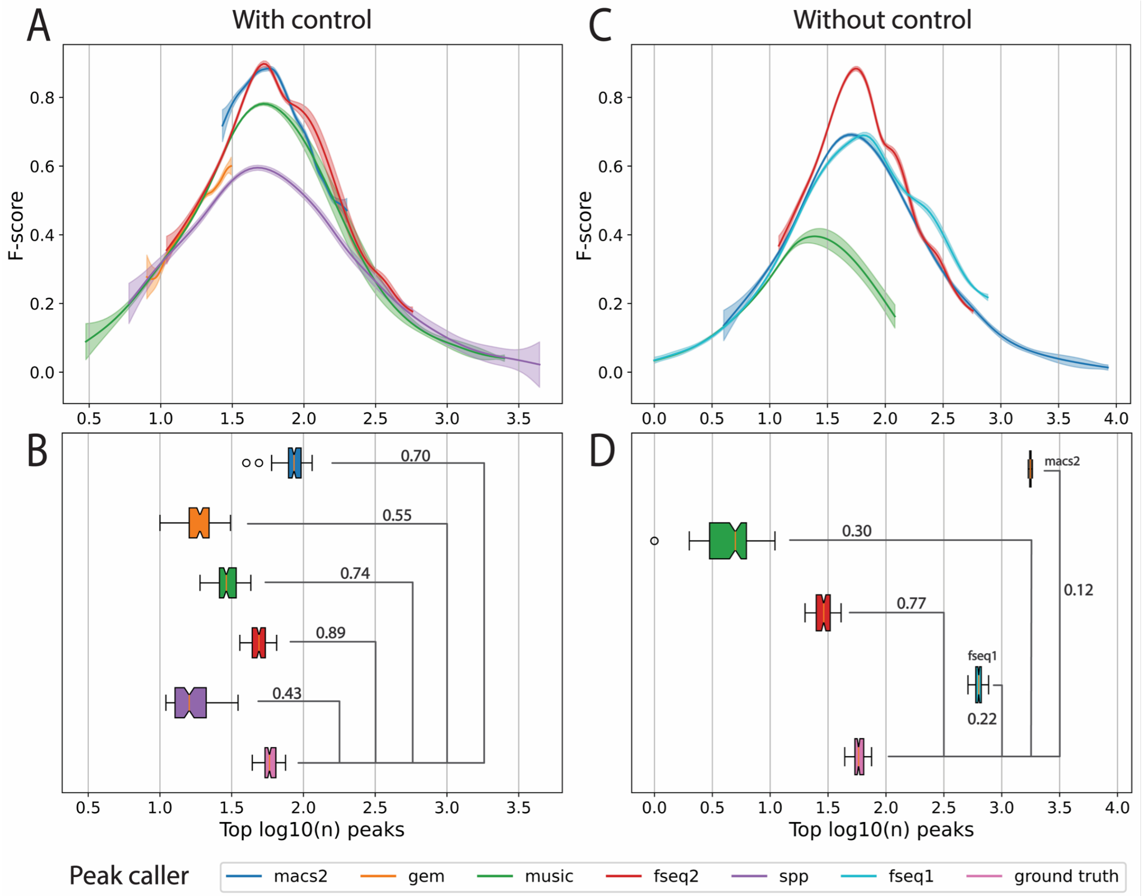


Figure 2.2: Comparison of peak callers on 100 pairs of simulated transcription factor ChIP-seq datasets. (A) The F-score operating characteristic curve where F-score is plotted as a function of the log10 top number of peaks called with control data. Generalized additive models are used to estimate the mean and 95% confidence intervals (shaded areas) of 100 peak calling results for each peak caller. (B) Boxplot of the number of peaks called by each peak caller with default threshold with control data, and the number of significant peaks in ground truth. Numbers are shown in log_{10} scale. Pearson's correlation coefficient r is shown above the bridge linking peak caller and ground truth. (C) The F-score plot without control data. SPP was not able to run without control. GEM resulted in few peaks which is not shown in the plot. (D) Boxplot without control data.

candidate summits, F-Seq2 demonstrated a highly balanced performance between precision and recall on simulated ChIP-seq data without controls (Fig 2.2C& D). F-Seq2 had the highest overall curve, which stood out among the other peak callers, including MACS2 and the original F-Seq, and achieved comparable performance (0.883) to those with control datasets (0.897). These results suggest that a significant amount of control information is contained within treatment dataset at a large scale. This is also evident in the real FoxA1 ChIP-seq dataset [21] where control read counts correlated well with treatment read counts in 10 kb windows across the genome. The observed high correlation and performance of F-seq2 implies that control information can be robustly extracted from treatment data and can be used to estimate background distribution for peak calling, given it does not greatly contradict with the treatment data and given a statistically rigorous modeling method for treatment data (e.g. F-Seq2 KDE). For real ChIP-seq datasets, especially where the correlation is low between control and treatment data, calling peaks without control data is less accurate due to the loss of unique information and cannot be recovered from treatment data [63].

2.4.2 Performance on real datasets

The absence of control data is more often seen in DNase-seq and ATAC-seq experiments compared to ChIP-seq. Therefore, F-seq2 was directly compared to MACS2 on ATAC-seq data to further evaluate performance in the absence of control data (Fig 2.3). Both F-Seq2 with paired-end (PE) auto mode and MACS2 with single-end (SE) shift-extend mode, which are two different strategies to avoid calling peaks on nucleosome centers, precisely identified open chromatin regions with their top ranked peaks (see Material and Methods for auto filter design details). The higher overall characteristic curve of F-Seq2 (highest F-0.5 score = 0.62) indicates the filter-based method is more effective in avoiding peaks called on nucleosomes compared to the shift-based method. MACS2 SE shift-extend mode outperformed its PE mode (highest F-0.5 scores: 0.58 vs. 0.54) at low genome coverage (1% of human genome). This precision gained by the shift-extend strategy is likely why single-end data is used as part of the official ENCODE ATAC-seq data analysis pipeline [5]. At larger genome coverage (2%), F-Seq2 PE without filter mode, and SE mode showed superior performance versus all other modes (both had the highest value for F-0.5 score = 0.62 at different coverages). This observation suggests that the additional data improved precision for medium ranked peaks in F-Seq2 in its non-filter-based mode, which takes advantage of the greater genomic information available for more robust and accurate background estimations at the cost of precision at low genome coverage.

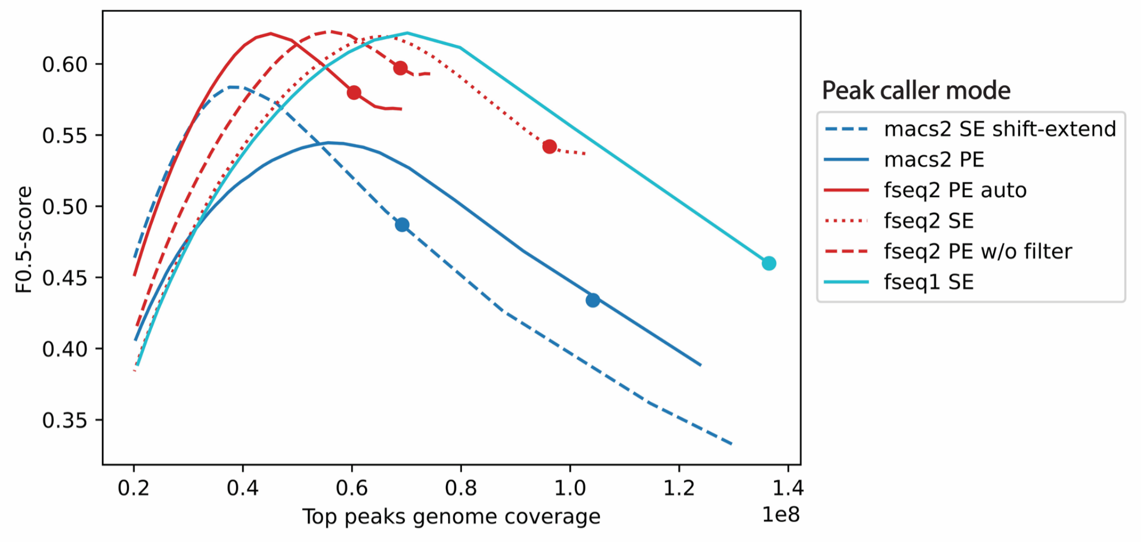


Figure 2.3: Comparison of F-Seq2 and MACS2 on the ATAC-seq paired-end data in GM12878. The F0.5-score operating characteristic curve where F0.5-score is plotted as a function of the genome coverage in base pairs by the top ranked peaks. F0.5-score put more emphasis on precision than recall due to the incompleteness of our "ground truth". MACS2 was run with two modes: SE shift-extend mode and PE mode. SE shift-extend mode first shifted both 5' and 3' ends 75 bp towards outside (5' end in 3' to 5' direction, 3' end in 5' to 3' direction), then extended 150 bp towards inside. This approach smoothed the counts of cutting events by the extension size, which is used by the ENCODE ATAC-seq data analysis pipeline [5]. F-Seq2 was run with three modes: PE auto mode, PE without filter mode, and SE mode. PE auto mode used the F-Seq2 auto filter which is designed based on nucleosome-related fragment length information (See Methods for design details). Dots on curves indicate the genome coverage of significant peaks by the default threshold of each peak caller.

Interestingly, the original F-Seq1 with SE mode had a similar characteristic curve to F-Seq2 with SE mode, and even better performance at larger genome coverage. The similar performance observed for both versions validates the assumption F-

Seq1 made that the peaks with higher signals are more likely to be true positives (versus false positives) in open chromatin datasets compared to those in ChIP-seq datasets. This alleviates the need to further conduct the dynamic Poisson tests in DNase-seq and ATAC-seq datasets while maintaining high F-0.5 scores. Despite the effectiveness of the dynamic Poisson test at filtering out false positives in ChIP-seq datasets, it potentially filters out more true positives in ATAC-seq datasets, shown by the superior performance of the original F-Seq with SE mode at larger genome coverage. F-Seq peak ranks can be reproduced in F-Seq2 by ranking peaks with signal values.

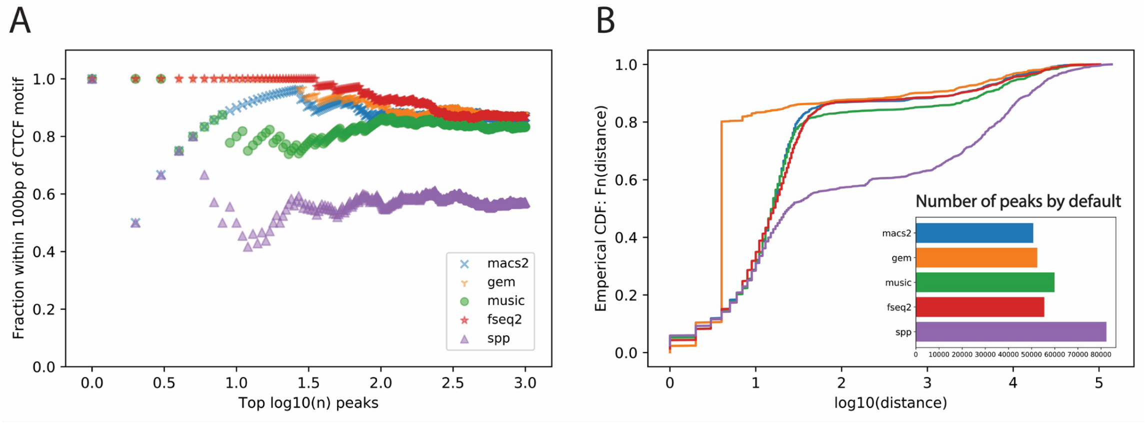


Figure 2.4: Comparison of peak callers on the CTCF ChIP-seq in ascending aorta female adult (51 years). (A) The fraction of top n peaks within 100 bp of a CTCF motif. (B) The empirical distribution of the shortest distance of the called peaks to a CTCF motif. The subplot shows the number of significant peaks called by each method using the default threshold.

F-Seq2 was benchmarked on 3 real ChIP-seq datasets to confirm that the observed high performance under the simulated situations can be recapitulated using real data. F-Seq2 had the largest fraction of top n peaks (up to 1000 peaks) within 100 bp of a CTCF motif (Fig 2.4). GEM was the second largest with slightly better performance than MACS2. The empirical distribution of the distance of called peaks to the nearest CTCF motif showed a clear performance advantage for GEM in detecting peaks centered around motifs: 80% of the 1000 most significant peaks were within 4 bp of a CTCF motif. This performance differential is due to GEM's utilization of motifs, where the tool intends to improve peak calling accuracy at the expense of increased run time, and potentially introducing bias by ranking peaks without motifs lower than those containing a TF motif. MACS2 and F-Seq2 had the shortest execution time for the CTCF datasets while maintaining favorable performance relative to GEM (Fig 2.5). Similar trends were observed in the MAFK ChIP-seq dataset benchmarking results, with SPP being an exception as it had the most variable number of peaks called by a default threshold between the two TFs (Fig 2.6). However, all peak callers had a much lower and barely distinguishable performance between each other on STAT1 (Fig 2.7). Karimzadeh and Hoffman [71] showed that 76 out of 220 chromatin factor ChIP-seq peaks lacked relevant sequence motifs, and STAT1 peaks were low in motif occupancy (below 50%), suggesting that evaluating peak callers using motifs may not reflect actual performance. As the motif-centered evaluation is likely problematic, it is necessary to use the more accurate and precise simulated ground truth data when assessing tool performance.

2.5 Discussion

The highly-balanced performance of F-Seq2 between precision and recall across different assays is noteworthy. Kernel density estimation (KDE), which is a nonparametric method to model the read probability distribution, has an advantage over explicit modeling methods. Confounding experimental and biological factors, such as antibody specificity, DNA susceptibility to enzymes, and sequencing read mappability, make it difficult to form explicit assumptions [29], especially across different assays. The advantage of KDE has been demonstrated by the original

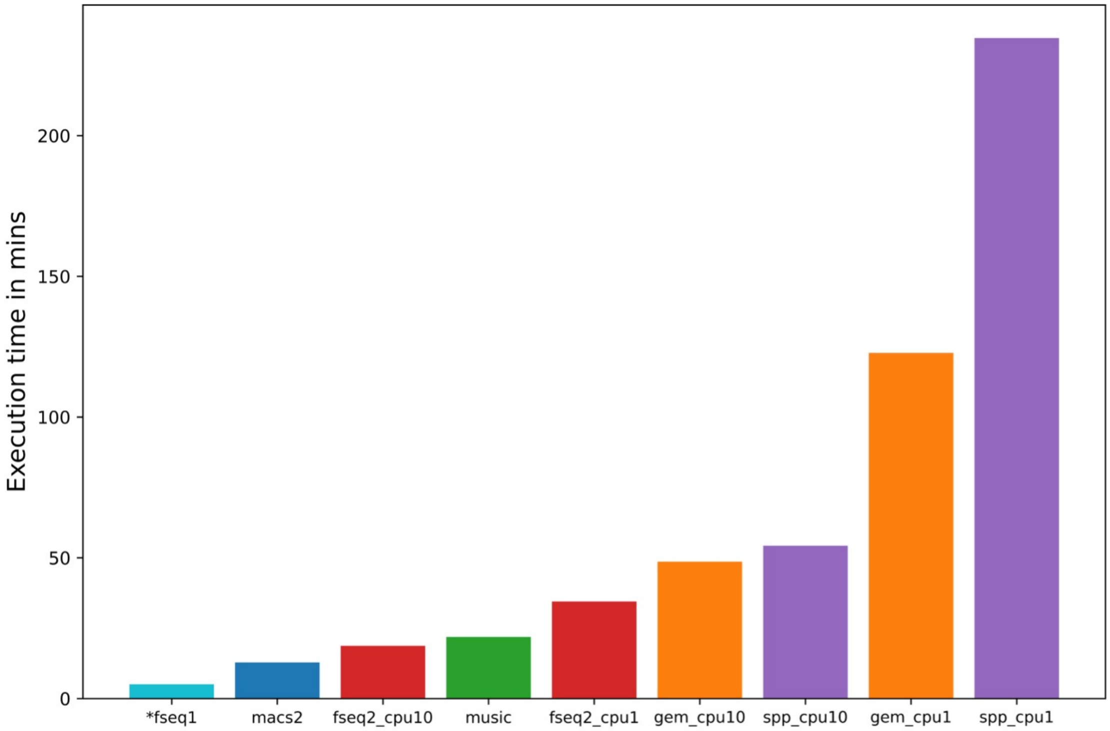


Figure 2.5: Execution time of peak callers on the CTCF ChIP-seq in ascending aorta female adult (51 years). F-Seq2, GEM, and SPP support and recommend multiprocessing; their performance with 10 CPU cores were included in the comparison. F-Seq1 was included as a baseline for comparison since the program did not accept bam files as input, nor support a control dataset (i.e. missing the bamtobed conversion time, and reading in and utilizing the control dataset time). Different sub-tasks were performed by peak callers, which affected the execution time. For example, GEM conducted the optional motif analysis to call peaks near the discovered motifs, and F-Seq2 reconstructed the genome-wide signal despite no output to a bigwig file.

peak caller F-Seq, which is the top-performing peak caller on DNase-seq datasets [7], and frequently used for FAIRE-seq data peak calling [57]. We designed a new statistical framework and introduced new features to F-Seq to further improve the performance in this second version. Adding support for user-input control data allows for F-Seq2 to more accurately model background reads distribution together with the treatment reads distribution. With the help of a dynamic parameter λ_{local} , read distributions around candidate summits can be summarized into significance values accounting for local biases, leading to statistically robust peak ranks and peak

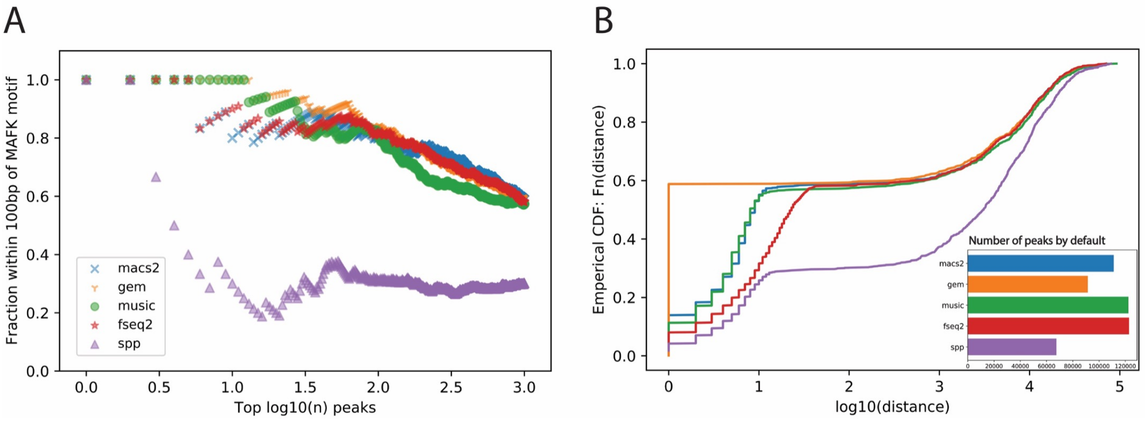


Figure 2.6: Comparison of peak callers on the MAFK ChIP-seq in HepG2. (A) The fraction of top n peaks within 100 bp of a MAFK motif. (B) The empirical distribution of the shortest distance of the called peaks to a MAFK motif. The subplot shows the number of significant peaks called by each method using the default threshold.

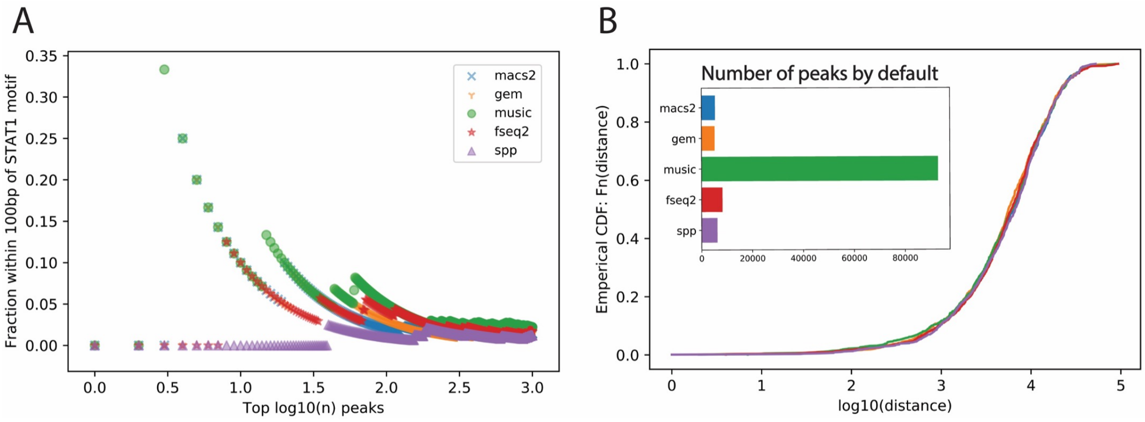


Figure 2.7: Comparison of peak callers on the STAT1 ChIP-seq in GM12878. (A) The fraction of top n peaks within 100 bp of a STAT1 motif. (B) The empirical distribution of the shortest distance of the called peaks to a STAT1 motif. The subplot shows the number of significant peaks called by each method using the default threshold.

calls. The joint effect of KDE and the dynamic parameter demonstrated superior performance in our benchmarking results, especially without control data. This suggests control information can be extracted from treatment data, given control and treatment data are well correlated. The support of control data allows for a more biologically meaningful signal to be reconstructed by weighting the treatment with control data, which leads to a better sanity-check when comparing and combining signals from different datasets [61].

Whether control data is a dispensable dataset for ChIP-seq peak calling requires further investigation. Recent papers [63, 72] that predict the linear weights for control datasets from treatment datasets provide evidence that control information can be extracted from treatment data. In our simulation results, a comparable performance was observed when using or omitting control data. F-Seq2 runs using experiments with real ChIP-seq data showed only a slightly decrease in performance without control data (data not shown). We suspect that the high correlations between control and treatment data explain the observation that control data is not required in a simulation setting. However, conclusions cannot be made based on the small performance difference on the real ChIP-seq datasets due to evaluation biases with motifs. We are unable to determine if a large observable discrepancy (low correlation) between control and treatment data is due to the low quality of either of the datasets, or to the indispensable information contained within control dataset.

F-Seq2 is compatible and suitable for IDR analysis which we recommend as a more reliable approach to determine a significance threshold when working with replicates. The IDR algorithm requires peak callers to run at a relaxed threshold to include both signal and noise peaks within the output to detect the consistency transition point between the two groups [64]. During benchmarking, the MACS2 peak width detection was observed to be tied to peak detection. When the q-value threshold was lowered, by default MACS2 called not only more peaks, but larger width peaks, and may cause irreproducibility as a side-effect (i.e. changing the significance scores and ranks of called peaks). We developed F-Seq2 with summit-focused statistical testing and used separate parameters for peak width detection and summit detection. F-Seq2 reliably reproduces the same exact summits and peaks when lowering the p-value or q-value threshold, and an individual significance score for each summit is calculated. Having separate scores for each summit and less rank ties by p-value interpolation are essential for IDR to precisely identify the transition point, representing the desired threshold. We have built a peak calling pipeline for a pair of replicates with F-Seq2 followed by an integrated IDR analysis with our recommended settings, which is directly accessible through the command line interface.

F-Seq2 further pushes the potential in the mature field of peak calling. The accuracy of peak calling is essential for downstream analysis, such as differential and motif analysis, to discover new biological insights and mechanisms with HTS data.

2.6 Availability

Data accessibility and peak caller parameter settings.

Simulated data was reproduced from Thomas *et al.* [32]. The adapted scripts to simulate ChIP-seq data, and the scripts to run all peak callers are available at `https://github.com/Boyle-Lab/F-Seq2-Paper-Supplementary`. The accession numbers of all ENCODE data, and the IDs of all JASPAR motifs used in this study are also available at this website.

Software availability.

The F-Seq2 software and documentation are available at `https://github.com/Boyle-Lab/F-Seq2`. F-Seq2 can be installed through the Python Package Index (PyPI) and the Conda package manager.

Supplementary Data are available at NAR online.

2.7 Publication

The study in this chapter has been published in *NAR Genomics and Bioinformatics* [73]: Zhao, N., & Boyle, A. P. (2021). F-Seq2: improving the feature density based peak caller with dynamic statistics. *NAR Genomics and Bioinformatics*, 3(1), lqab012.

CHAPTER III

Creating a Comprehensive and Accessible Database of Human Non-coding Regulatory Variants

3.1 Introduction

Nearly 90% of the disease risk-associated variants identified from genome-wide association studies (GWAS) are in non-coding regions of the genome. The annotations obtained from analyzing functional genomics assays can provide additional information to pinpoint causal variants, which are often not the lead variants identified from association studies. However, the lack of available annotation tools limits the use of such data. To address the challenge, we have previously built the RegulomeDB database for prioritizing and annotating variants in non-coding regions [74], which has been a highly utilized resource for the research community (Fig 3.1).

Here we present an update of the RegulomeDB web server, RegulomeDB v2 (`http://regulomedb.org`). RegulomeDB annotates a variant by intersecting its position with genomic intervals identified from functional genomic assays and computational approaches. It also incorporates those hits of a variant into a heuristic ranking score, representing its potential to be functional in regulatory elements. We improve and boost annotation power by incorporating thousands of newly processed data from functional genomic assays in GRCh38 assembly, and include probabilistic scores from the SURF algorithm that was the top-performing non-coding variant

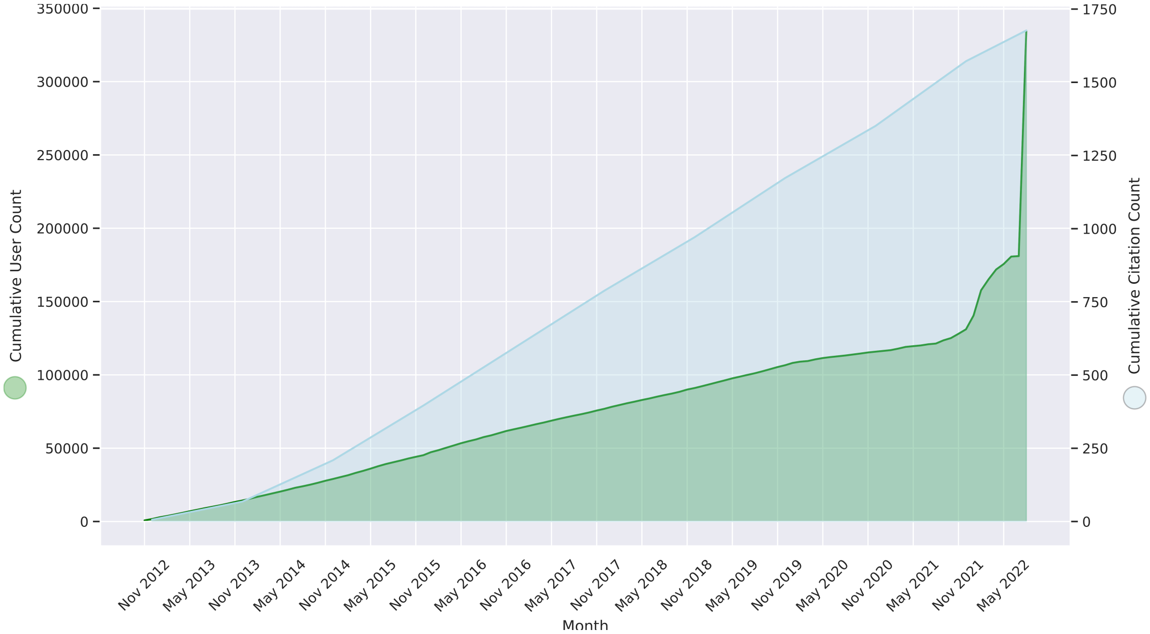


Figure 3.1: Popularity of RegulomeDB. The x-axis is month and year since RegulomeDB first published in 2012. The left y-axis is cumulative user count (green). The right y-axis is cumulative citation count (light blue). The citation count data are derived from Clarivate Web of Science. © Copyright Clarivate 2022. All rights reserved.

predictor in CAGI 5 [75].

3.2 Methods

3.2.1 Data sources

Genomic variants

The information of genomic variants was retrieved from dbSNP153 [76], including the positions and allele frequencies from different projects, such as the 1000 genome project [77], TOPMED [78], and GnomAD [79].

ChIP-seq and chromatin accessibility experiments

We collected the peaks of ChIP-seq targeting transcription factors (TF), DNase-seq, and ATAC-seq experiments called by uniform pipeline from the latest release of the ENCODE portal, which includes the experiments from the Roadmap project [80].

PWM matching

We downloaded the PWMs (position weight matrices) of 746 non-redundant TF motifs from the JASPAR 2020 database [81]. The kmers matching to TF motifs were called by TFM P-value with a threshold at 4^{-8} for each PWM [82]. Bowtie was used to map the kmers on the genome to determine the final PWM matching positions for the TF motifs [83]. The information content from each PWM was also integrated into the database and used as a feature to calculate the probabilistic score from the random forest model.

Footprints

Footprints were predicted with signals from 642 DNase-seq experiments and 591 TF motifs by the TRACE pipeline: https://www.encodeproject.org/search/?type=Annotation&internal_tags=RegulomeDB_2_2&annotation_type=footprints&software_used.software.name=trace [84]. TRACE is a computational method that incorporates DNase-seq signals and PWMs within a multivariate hidden Markov model to detect footprint regions with matching motifs.

Chromatin states

Chromatin states in 833 biosamples were called from chromHMM in EpiMap [85] and were directly retrieved from the ENCODE portal.

eQTLs

The eQTLs from the GTEx project across 49 human tissues were downloaded from the GTEx portal (https://storage.googleapis.com/gtex_analysis_v8/single_tissue_qtl_data/GTEx_Analysis_v8_eQTL.tar) [86]. The variant-gene pairs with the corresponding tissue were added as annotations in the database.

caQTLs

The chromatin accessibility QTLs (caQTLs) were collected from 9 publications [87, 88, 89, 90, 91, 92, 93, 94, 95] https://www.encodeproject.org/search/?type=A `nnotation&internal_tags=RegulomeDB_2_2&annotation_type=caQTLs`. Only SNVs were included and lifted over from hg19 to GRCh38 if necessary [96].

Prediction scores

We provide a heuristic ranking and a probabilistic score for each query variant representing its potential of being a functional variant in regulatory elements. The heuristic ranking is defined in the same way as in the previous version of RegulomeDB [74]. The probabilistic score is calculated from a random forest model, TURF, trained with allele-specific TF binding SNVs [97]. We used a simplified version here only including binary features from functional genomic evidence as used in the heuristic ranking, as well as numeric features from information content in matched PWMs. We will include the whole feature set in a future release.

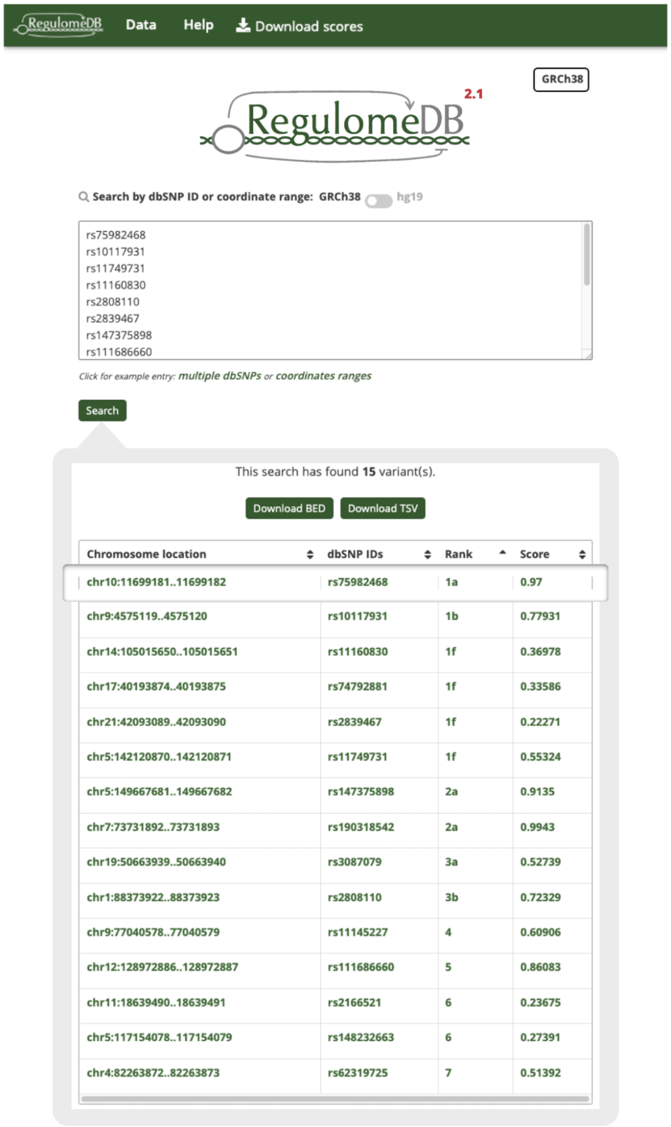
3.2.2 Database and web server design

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 the ENCODE portal creating a genomic data service (`https://github.org/ENCODE-DCC/genomic-data-service`). The genomic intervals are parsed from BED formatted files and associated with metadata of the source experiments and computational pipelines from the ENCODE portal. These BED files are then indexed in Elasticsearch (`https://www.elastic.co/`) as in integer range type to enable efficient search against a query position. In total, over two billion genomic intervals representing ChIP-seq and DNase-seq peaks, matches to PWMs and DNase footprints, eQTLS, caQTLs and chromatin states 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 TURF [75, 97]. The query results are displayed with a web interface (`https://github.org/ENCODE-DCC/regulome-encoded`) that contains charts and interaction figures, which can be customized by users.

3.2.3 New interface for variant functionality exploration

The RegulomeDB v2 web server accepts any query variant on the whole genome in either GRCh38 or hg19 genome assembly. A toggle above the search box allows users to switch between the two assemblies. The search box allows any user to input multiple queries (up to 500 at a time) (Fig 3.2). The input query can be in three formats: 1) rsID (from dbSNP database v153); 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 more than 1% allele frequency from dbSNP153 will be queried. The backend then intersects the variant(s) position with the genomic intervals of annotations obtained predicted from functional genomics experiments and returns a sortable summary table of variant scores (Fig 3.2), including a ranking score and a probabilistic score showing its potential of being a regulatory variant. In addition, a dbSNP rsID will link to the query variant if it exists.

After 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 (Fig 3.3, Fig 3.4). The top of the page shows some basic information on the variant position, scores, and allele frequencies from the dbSNP database. While on the bottom is the initial summary section on genomic annotations' hits. Since a single query can hit up to 2,000 results, the initial summary section is divided into five data types; TF binding sites from ChIP-seq, chromatin states from chromHMM, chromatin accessibility, PWM matching or footprint predictions, and eQTLs or caQTLs. In addition, a genome



Search by dbSNP ID or coordinate range: GRCh38 ○ hg19

```
rs75982468
rs10117931
rs11749731
rs11160830
rs2808110
rs2839467
rs147375898
rs111686660
```

Click for example entry: *multiple dbSNPs* *or* *coordinates ranges*

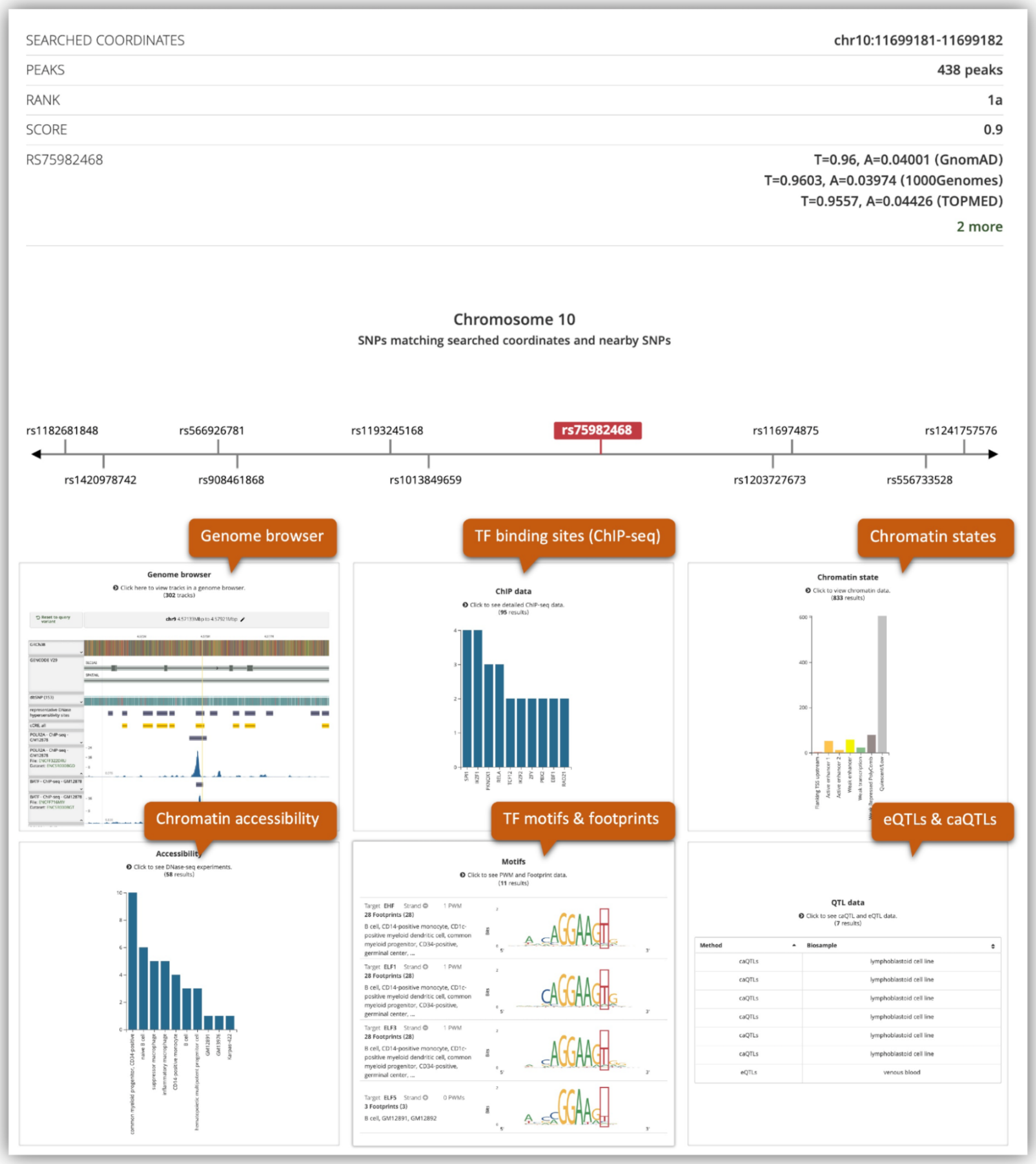
This search has found **15** variant(s).

Chromosome location	dbSNP IDs	Rank	Score
chr10:11699181..11699182	rs75982468	1a	0.97
chr9:4575119..4575120	rs10117931	1b	0.77931
chr14:105015650..105015651	rs11160830	1f	0.36978
chr17:40193874..40193875	rs74792881	1f	0.33586
chr21:42093089..42093090	rs2839467	1f	0.22271
chr5:142120870..142120871	rs11749731	1f	0.55324
chr5:149667681..149667682	rs147375898	2a	0.9135
chr7:73731892..73731893	rs190318542	2a	0.9943
chr19:50663939..50663940	rs3087079	3a	0.52739
chr1:88373922..88373923	rs2808110	3b	0.72329
chr9:77040578..77040579	rs11145227	4	0.60906
chr12:128972886..128972887	rs111686660	5	0.86083
chr11:18639490..18639491	rs2166521	6	0.23675
chr5:117154078..117154079	rs148232663	6	0.27391
chr4:82263872..82263873	rs62319725	7	0.51392

Figure 3.2: RegulomeDB Query Interface. An example query with the rsIDs of variants from dbSNP database. Upon clicking the search buttons, a summary table representing prediction scores for all query variants will be displayed.

browser section is also available to view the specific DNase-seq and ChIP-seq data, which can aid in 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. The chromatin state tab shows the chromHMM state for each of the 833 biosamples, which also includes an intuitive body map colored by the



SEARCHED COORDINATES	chr10:11699181-11699182
PEAKS	438 peaks
RANK	1a
SCORE	0.9

RS75982468	T=0.96, A=0.04001 (GnomAD)
	T=0.9603, A=0.03974 (1000Genomes)
	T=0.9557, A=0.04426 (TOPMED)
	2 more

Chromosome 10
SNPs matching searched coordinates and nearby SNPs

Genome browser
Click here to view tracks in a genome browser.
(302 tracks)

ChIP data
Click to see detailed ChIP-seq data.
(95 results)

Chromatin state
Click to view chromatin data.
(833 results)

Accessibility
Click to see DNase-seq experiments.
(58 results)

Motifs
Click to see PWM and Footprint data.
(11 results)

Target EHF Strand ⊖ 1 PWM
28 Footprints (28)
B cell, CD14-positive monocyte, CD1c-positive myeloid dendritic cell, common myeloid progenitor, CD34-positive, germinal center, ...

Target ELF1 Strand ⊕ 1 PWM
28 Footprints (28)
B cell, CD14-positive monocyte, CD1c-positive myeloid dendritic cell, common myeloid progenitor, CD34-positive, germinal center, ...

Target ELF3 Strand ⊖ 1 PWM
28 Footprints (28)
B cell, CD14-positive monocyte, CD1c-positive myeloid dendritic cell, common myeloid progenitor, CD34-positive, germinal center, ...

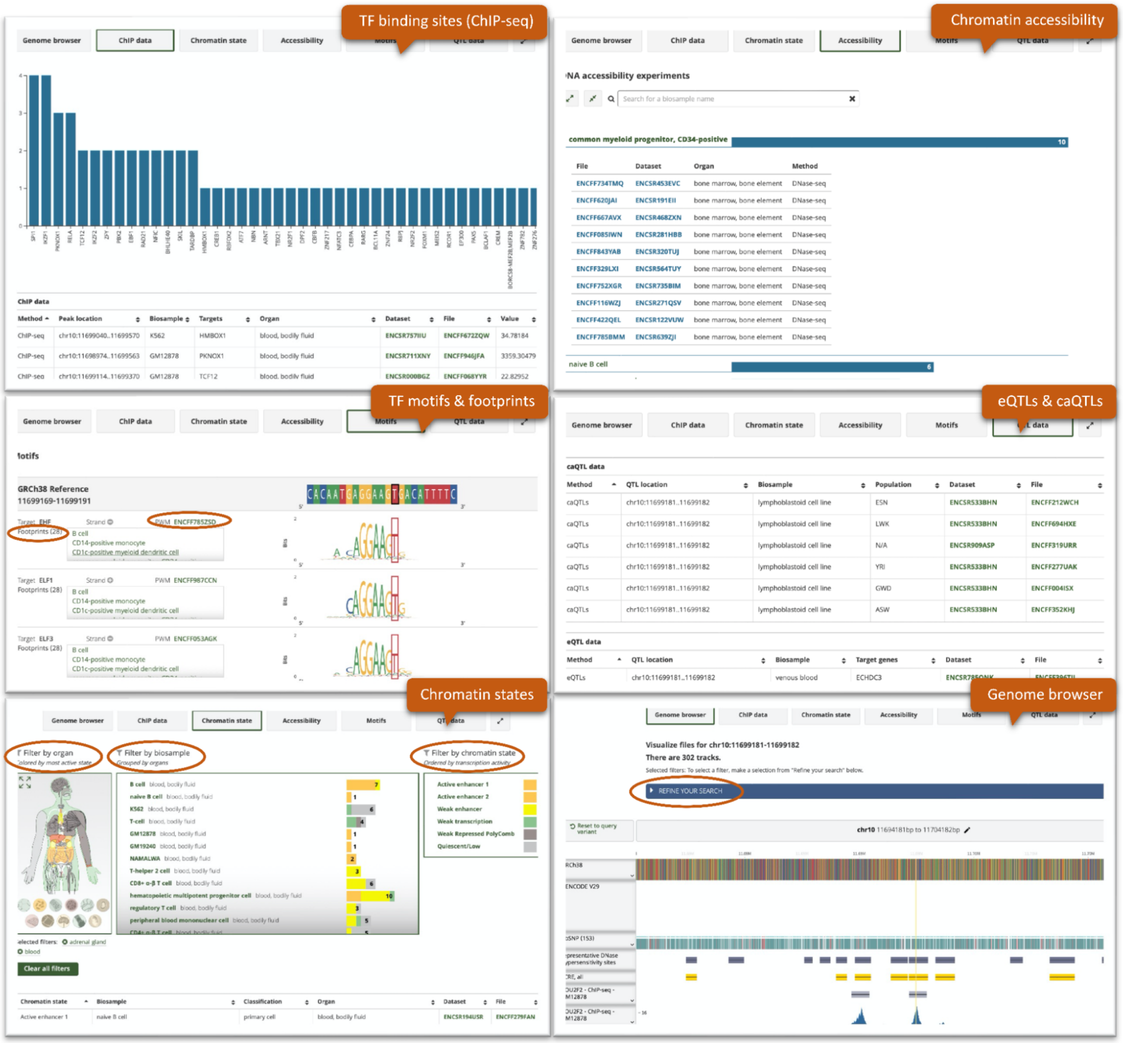
Target ELF5 Strand ⊕ 0 PWMs
3 Footprints (3)
B cell, GM12891, GM12892

QTL data
Click to see caQTL and eQTL data.
(7 results)

Method	Biosample
caQTLs	lymphoblastoid cell line
caQTLs	lymphoblastoid cell line
caQTLs	lymphoblastoid cell line
caQTLs	lymphoblastoid cell line
caQTLs	lymphoblastoid cell line
caQTLs	lymphoblastoid cell line
eQTLs	venous blood

Figure 3.3: RegulomeDB Result Overview Page of rs75982468. For any variant of interest, a results page with more information on the hits of genomic annotations is available. Each of the six sections at the bottom can be clicked to expand more details on each data type.

most active chromatin state in each organ. 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 (shown as a yellow highlight). The tracks on the genome browser can be further filtered using a modal that allows one



ChIP data

Method	Peak location	Biosample	Targets	Organ	Dataset	File	Value
ChIP-seq	chr10:11699040..11699570	K562	HMBOX1	blood, bodily fluid	ENCSR757IIU	ENCFF672ZQW	34.78184
ChIP-seq	chr10:11698974..11699563	GM12878	PKNOX1	blood, bodily fluid	ENCSR711XNY	ENCFF946JFA	3359.30479
ChIP-seq	chr10:11699114..11699370	GM12878	TCF12	blood, bodily fluid	ENCSR000BGZ	ENCFF068YYR	22.82952

NA accessibility experiments

Search for a biosample name

common myeloid progenitor, CD34-positive 10

File	Dataset	Organ	Method
ENCFF734TMQ	ENCSR453EVC	bone marrow, bone element	DNase-seq
ENCFF620JAI	ENCSR191EII	bone marrow, bone element	DNase-seq
ENCFF667AVX	ENCSR468ZXN	bone marrow, bone element	DNase-seq
ENCFF085IWN	ENCSR281HBB	bone marrow, bone element	DNase-seq
ENCFF843YAB	ENCSR320TUJ	bone marrow, bone element	DNase-seq
ENCFF329LXI	ENCSR564TUY	bone marrow, bone element	DNase-seq
ENCFF752XGR	ENCSR735BIM	bone marrow, bone element	DNase-seq
ENCFF116WZJ	ENCSR271QSV	bone marrow, bone element	DNase-seq
ENCFF422QEL	ENCSR122VUW	bone marrow, bone element	DNase-seq
ENCFF785BMM	ENCSR639ZJI	bone marrow, bone element	DNase-seq

naive B cell 6

Motifs

GRCh38 Reference
11699169-11699191

Target EHF Strand ⊖ PWM ENCFF785ZSD
Footprints (28)
B cell
CD14-positive monocyte
CD1c-positive myeloid dendritic cell

Target ELF1 Strand ⊕ PWM ENCFF987CCN
Footprints (28)
B cell
CD14-positive monocyte
CD1c-positive myeloid dendritic cell

Target ELF3 Strand ⊖ PWM ENCFF053AGK
Footprints (28)
B cell
CD14-positive monocyte
CD1c-positive myeloid dendritic cell

caQTL data

Method	QTL location	Biosample	Population	Dataset	File
caQTLs	chr10:11699181..11699182	lymphoblastoid cell line	ESN	ENCSR533BHN	ENCFF212WCH
caQTLs	chr10:11699181..11699182	lymphoblastoid cell line	LWK	ENCSR533BHN	ENCFF694HXE
caQTLs	chr10:11699181..11699182	lymphoblastoid cell line	N/A	ENCSR909ASP	ENCFF319URR
caQTLs	chr10:11699181..11699182	lymphoblastoid cell line	YRI	ENCSR533BHN	ENCFF277UAK
caQTLs	chr10:11699181..11699182	lymphoblastoid cell line	GWD	ENCSR533BHN	ENCFF004ISX
caQTLs	chr10:11699181..11699182	lymphoblastoid cell line	ASW	ENCSR533BHN	ENCFF352KHJ

eQTL data

Method	QTL location	Biosample	Target genes	Dataset	File
eQTLs	chr10:11699181..11699182	venous blood	ECHDC3	ENCSR785QNK	ENCFF396TJI

Biosample	Organ	Count
B cell	blood, bodily fluid	7
naive B cell	blood, bodily fluid	1
K562	blood, bodily fluid	6
T-cell	blood, bodily fluid	4
GM12878	blood, bodily fluid	1
GM19240	blood, bodily fluid	1
NAMALWA	blood, bodily fluid	2
T-helper 2 cell	blood, bodily fluid	3
CD8+ α-β T cell	blood, bodily fluid	6
hematopoietic multipotent progenitor cell	blood, bodily fluid	10
regulatory T cell	blood, bodily fluid	3
peripheral blood mononuclear cell	blood, bodily fluid	5
CD4+ α-β T cell	blood, bodily fluid	5

Chromatin state	
Active enhancer 1	🟧
Active enhancer 2	🟧
Weak enhancer	🟨
Weak transcription	🟩
Weak Repressed PolyComb	🟫
Quiescent/Low	⬜

Selected filters: adrenal gland blood

Chromatin state	Biosample	Classification	Organ	Dataset	File
Active enhancer 1	naive B cell	primary cell	blood, bodily fluid	ENCSR194USR	ENCFF279FAN

Visualize files for chr10:11699181-11699182
There are 302 tracks.
Selected filters: To select a filter, make a selection from "Refine your search" below.

chr10 11694181bp to 11704182bp

Figure 3.4: RegulomeDB Expanded Pages of rs75982468. The expanded pages of each section shows details on the genomic experiments and annotations, such as the biosample, organ, TF target and the peak file called from the ENCODE project. The body map under the chromatin states view is colored by the most active state among all biosamples in each organ, which gives an intuitive way to explore the candidate organs where the query variant might be functional. Users can also explore the nearby genes of the query variant under the genome browser view.

to sub-select by specific organ/cell types, biosample types, file types, assay methods, or by TF targets.

3.3 Results

The update of RegulomeDB now includes >650 million and >1.5 billion genomic intervals in hg19 and GRCh38, respectively, a 5x increase over the previous version (Fig 3.5). We included approximately 5,000 transcription factor (TF) ChIP-seq and chromatin accessibility experiments from the ENCODE project [5, 80], the Roadmap Epigenomics Consortium [5, 80], and the Genomics of Gene Regulation Consortium. We also produced a comprehensive set of footprint predictions using over 800 chromatin accessibility experiments and 591 TF motifs in GRCh38 using the TRACE pipeline [84]. In addition, we refined the included TF motifs by using the non-redundant vertebrates set from the JASPAR database [81]. We also integrated approximately 71 million variant-gene pairs in eQTL studies from the GTEx project [86], and 450,000 caQTLs (chromatin accessibility QTLs) from 9 recent publications. Finally, we included chromatin state annotations called from chromHMM in EpiMap for 833 biosamples [85].

RegulomeDB accepts any query variants genome-wide in either GRCh38 or hg19 genome assembly by rsID or genome coordinates. The query variants can then be prioritized by functional prediction scores shown in a sortable table. For any variant of interest, an information page on five types of supported genomic evidence, as well as a genome browser view is displayed. Each of the six sections can be clicked to show more detail for functionality exploration (Fig 3.2, Fig 3.3, Fig 3.4).

RegulomeDB enables researchers to quickly separate functional variants from a large pool of variants and assign tissue or organ specificity for each variant. Here we showcase this using four verified variants from recent literature [98, 99, 100, 101, 102], and demonstrate the applicability of RegulomeDB to annotate those variants based

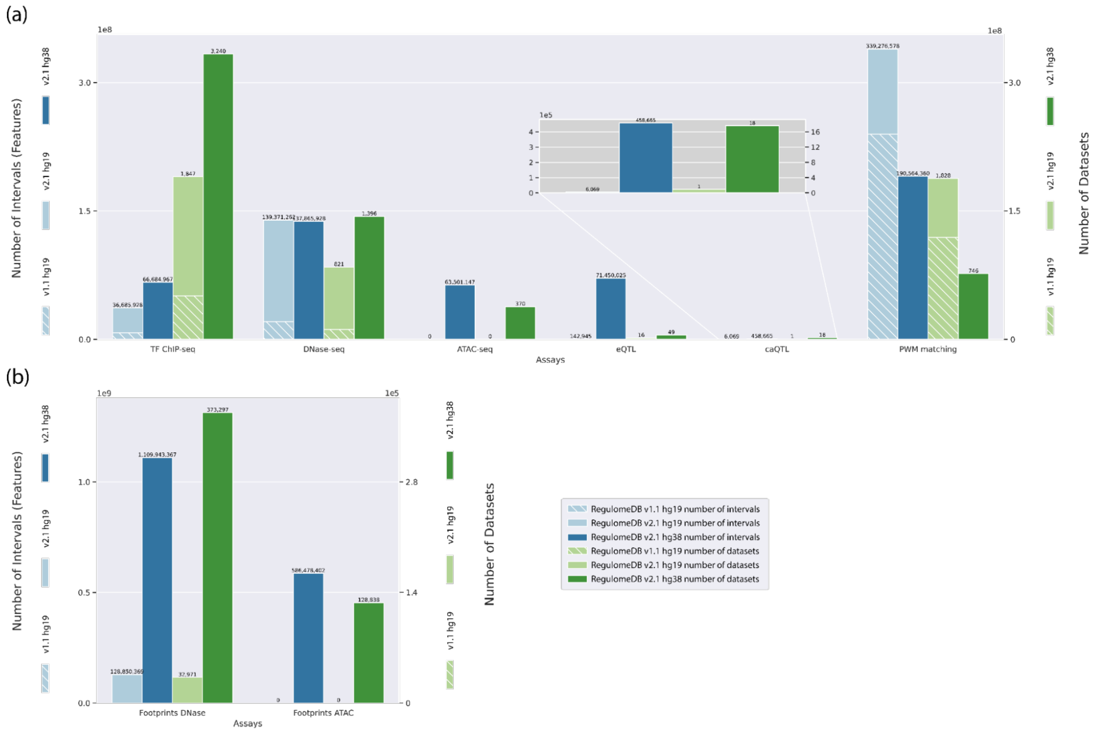


Figure 3.5: Overview of RegulomeDB Version 2 Data Growth and Refinement. Statistics on database content. Numbers under each data type include all experiments across different treatment conditions and biosamples. All numbers are RegulomeDB v2 stats, in hg19 or hg38.

on various sources of data (Fig 3.6).

TF motifs and ChIP-seq data together provide evidence about how a variant is likely to affect phenotype in a cell-specific context. For example, rs213641 is known to affect behavioral responses to fear and anxiety stimuli [98]. The POLR2A binding and the active transcriptional start site (TSS) state in the brain indicate that rs213641 is likely to function in the brain through disrupting the TSS of STMN1. We also examined rs7789585 where RegulomeDB TF motif evidence suggests that mutation to the reference allele G would disrupt the binding of GCM1, which may interrupt the active enhancer state at the locus in the heart. Hocker and colleagues recently confirmed this hypothesis using reporter assays and discovered that rs7789585

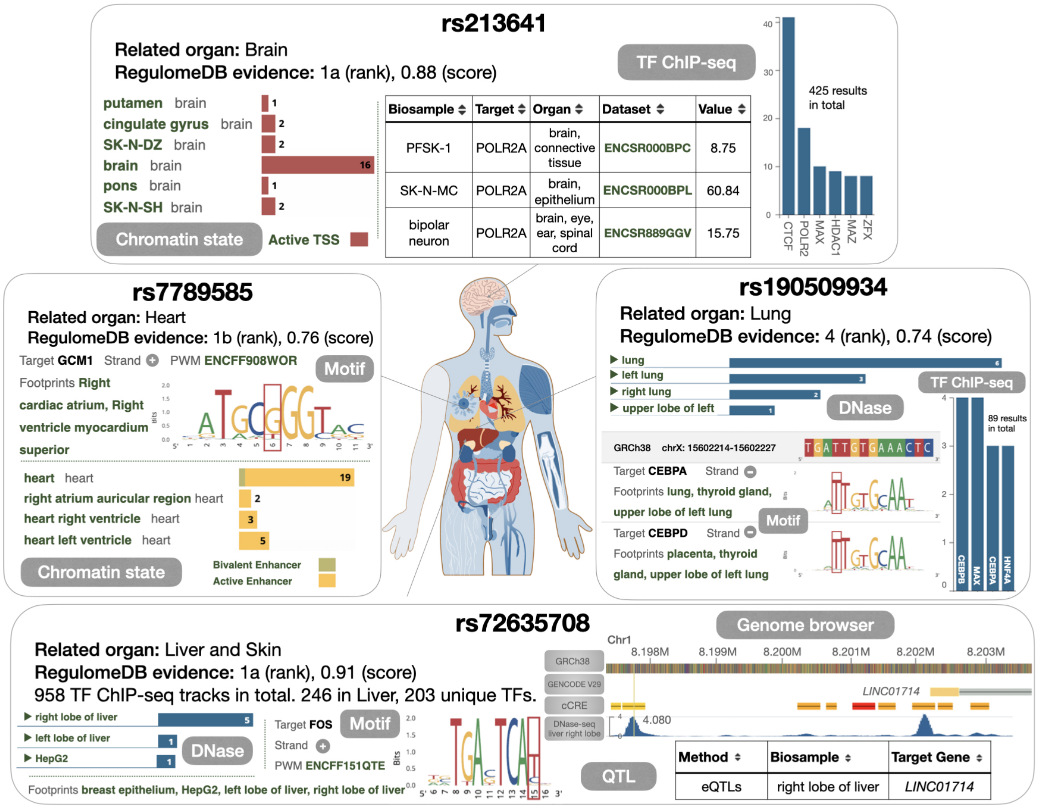

rs213641
Related organ: Brain
RegulomeDB evidence: 1a (rank), 0.88 (score)
TF ChIP-seq
putamen brain 1
cingulate gyrus brain 2
SK-N-DZ brain 2
brain brain 16
pons brain 1
SK-N-SH brain 2
Chromatin state
Active TSS
Biosample Target Organ Dataset Value
PFSK-1 POLR2A brain, connective tissue ENCSR000BPC 8.75
SK-N-MC POLR2A brain, epithelium ENCSR000BPL 60.84
bipolar neuron POLR2A brain, eye, ear, spinal cord ENCSR889GGV 15.75
425 results in total
CTCF
POLR2
MAX
HDAC1
MAZ
ZFX
rs7789585
Related organ: Heart
RegulomeDB evidence: 1b (rank), 0.76 (score)
Target GCM1 Strand + PWM ENCFF908WOR
Motif
Footprints Right cardiac atrium, Right ventricle myocardium superior
heart heart 19
right atrium auricular region heart 2
heart right ventricle heart 3
heart left ventricle heart 5
Chromatin state
Bivalent Enhancer
Active Enhancer
rs190509934
Related organ: Lung
RegulomeDB evidence: 4 (rank), 0.74 (score)
lung
left lung
right lung
upper lobe of left
TF ChIP-seq
DNase
89 results in total
GRCh38 chrX: 15602214-15602227
TGATTGTGAAACTC
Target CEBPA Strand −
Footprints lung, thyroid gland, upper lobe of left lung
Motif
Target CEBPD Strand −
Footprints placenta, thyroid gland, upper lobe of left lung
CEBPB
MAX
CEBPA
HNF4A
rs72635708
Related organ: Liver and Skin
RegulomeDB evidence: 1a (rank), 0.91 (score)
958 TF ChIP-seq tracks in total. 246 in Liver, 203 unique TFs.
Genome browser
right lobe of liver 5
left lobe of liver 1
HepG2 1
DNase
Target FOS Strand + PWM ENCFF151QTE
Motif
Footprints breast epithelium, HepG2, left lobe of liver, right lobe of liver
Chr1
8.198M
8.199M
8.200M
8.201M
8.202M
8.203M
GRCh38
GENCODE V29
cCRE
DNase-seq liver right lobe
LINC01714
4.080
QTL
Method Biosample Target Gene
eQTLs right lobe of liver LINC01714

Figure 3.6: Prioritization of Functional Variants with RegulomeDB Version 2. Four example variants with verified functions in related organs from recent literature. Various sources of evidence in RegulomeDB are indicated by gray boxes. RegulomeDB heuristic ranking score and probability score summarized all evidence.

disrupts a KCNH2 enhancer and affects cardiomyocyte electrophysiologic function [99].

DNase-seq assays and underlying footprint predictions identify open chromatin regions with mapped TF binding sites in hundreds of biosamples and can also be used to assign putative function to variants. rs190509934 has been associated with COVID-19 infection risk by affecting ACE2 expression level [100]. RegulomeDB shows hits to multiple DNase-seq peaks in lung-related biosamples. Furthermore, RegulomeDB extends this tissue effect with the hypothesis that ACE2 expression level may be regulated by CEBP by its overlap with DNase footprints in the lung found in the upstream promoter region of ACE2 [101]. In addition, eQTL studies provide correlation evidence between the variants and their target genes. For example, rs72635708 is predicted as a regulatory variant by RegulomeDB with a high probability of 0.91 due to its locus overlapping with DNase and ChIP-seq peaks, and footprints, and it is an eQTL associating with LINC01714 gene expression level in the right lobe liver. Since rs72635708 lies in the FOS motif, it is likely to be a functional variant in the liver by modulating the AP-1 complex binding level [102].

3.4 Summary

RegulomeDB provides a user-friendly tool to annotate and prioritize variants in non-coding regions of the human genome, which can aid variant function interpretation and guide follow-up experiments. We welcome user feedback through `regulomedb@mailman.stanford.edu`.

3.5 Publication

The study in this chapter has been published in *Nature Genetics* [103]: Dong, S.*, Zhao, N.*, Spragins, E., Kagda, M. S., Li, M., Assis, P. R., ... & Hitz, B. C. (2023). Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nature Genetics*, 2023-04.

CHAPTER IV

Prioritization of Regulatory Variants with Organ-specific Function in Non-coding Regions of the Human Genome

4.1 Abstract

Identifying non-coding regulatory variants in the human genome remains a challenging task in genomics. Recently we advanced our leading regulatory variants database, RegulomeDB, to its second version. Building upon this comprehensive database, we developed a novel machine-learning architecture with stacked generalization, TLand, which utilizes RegulomeDB-derived features to predict regulatory variants at cell or organ-specific levels. In our holdout benchmarking, TLand consistently outperformed state-of-the-art models, demonstrating its ability to generalize to new cell lines or organs. We trained three types of organ-specific TLand models to overcome the common model bias toward high data availability cell lines or organs. These models accurately prioritize relevant organs for 2 million GWAS SNPs associated with GWAS traits. Moreover, our analysis of top-scoring variants in specific organ models showed a high enrichment of relevant GWAS traits. We expect that TLand and RegulomeDB will further advance our ability to understand human regulatory variants genome-wide.

4.2 Introduction

Understanding the biological impact of variants located in non-coding regions of the human genome is a significant challenge. Nearly 90% of disease risk-associated single nucleotide polymorphisms (SNPs) identified from genome-wide association studies (GWAS) are within non-coding regions. Similarly, 75% of patients affected by Mendelian disease have mutations outside of protein-coding regions [104]. The abundance of disease-associated non-coding variants makes studies focused on these regions highly desirable and thus facilitates understanding their functional consequences.

Prioritizing non-coding variants requires integrating multiple layers of functional information, including regulatory annotations identified from high-throughput sequencing datasets (e.g. DNase-seq [105], ChIP-seq [106], and ATAC-seq [8]). Such annotations provide additional information in pinpointing causal variants, which are often not the lead variants identified in GWAS studies. Despite the benefit of incorporating functional genomics assay-based evidence when examining non-coding variants, the lack of available annotation tools limits the use of such data. The majority of resources developed for clinical purposes have focused on coding regions as an application of exome sequencing-based data [107, 108], which captures less than 5% of human variation [109, 110, 77].

Previously we built RegulomeDB, a comprehensive database for prioritizing and annotating variants in non-coding regions, which was highly sought after in the research community [36]. RegulomeDB intersects query variants with regulatory regions predicted by functional genomics assays and, by utilizing ranking heuristics, informs users about putative functional consequences to prioritize variants. Recently, RegulomeDB has been upgraded to v2, to improve its annotation power by incorporating thousands of new functional genomics assays from the ENCODE project [5], Roadmap Epigenomics Consortium [80], and the Genomics of Gene Regulation Consortium. A suite of models, namely SURF and TURF, was developed and integrated in this version to provide accurate probabilistic scores for general and cell-type specific regulatory activities [38, 39].

However, a current drawback to this model suite is it was only optimized to predict function in six common cell lines from ENCODE using only the hg19-referenced datasets available at the time. The resulting models have a bias and lack sufficient statistical power to make generalized predictions toward less-studied cell lines. As a result, the RegulomeDB scores are less informative for cell lines, tissues, and organs that do not have the abundance of data available for the commonly studied cell lines. This can make it challenging for RegulomeDB users to identify targets of interest when screening variants and creating hypotheses regarding their regulatory functions in less-studied cell lines or organs. Nonetheless, we anticipate RegulomeDB to further grow for years to come as the ENCODE and the IGVF Consortium [111] repository incorporate more datasets, cells, tissues, and assay types, e.g., high-dimensional experimental assays like Hi-C [112] and computational annotations like Enformer [52]. However, the current SURF and TURF models were not designed to incorporate all of these data efficiently, and continuously, nor can they combat overfitting due to the expanding feature space.

Here we present TLand, a flexible architecture based on stacked generalization [55] to learn RegulomeDB-derived features to predict regulatory variants at a cell-specific level or organ-specific level. TLand took advantage of features derived from RegulomeDB v2, which now incorporates >650 million and >1.5 billion genomic intervals in hg19 and GRCh38, respectively, a 5x increase over the previous version [103]. Additionally, recently developed deep learning model predictions from Sei [47], the successor of DeepSEA, along with 1372 newly quantile normalized DNase signals were introduced to the TLand feature space. TLand's stacked generalization approach groups feature into biologically meaningful subspaces, training individual estimators before assembling them to reduce overfitting and enable further integration of features. Cell-specific TLand consistently outperformed state-of-the-art models during benchmarking in the hold-out cell line. Organ-specific TLand further improved upon the cell-specific TLand models in predicting organ-specific regulatory variants, and accurately prioritized relevant organs for GWAS traits by addressing the data availability bias by developing a suite of models. Furthermore, analysis of top-scoring variants in specific organ models showed high enrichment of correlated GWAS traits. Given its superior performance relative to its predecessors and competing methods, we expect TLAND to address the ongoing challenge of reliably prioritizing variants, even in less-studied cell lines and organs, thus advancing our ability to identify regulatory variants genome-wide.

4.3 Methods

4.3.1 Allele-specific binding (ASB) variants

We define a variant that exhibits regulatory function if it affects any TF-binding regulatory activities. We trained our models on the allele-specific TF binding (ASB) variants, which are defined as variants that result in significantly different TF binding affinities between two alleles at heterozygous sites in an individual. We included a total of 7,530 ASB variants in 6 cell lines (GM12878, HepG2, A549, K562, MCF7, and H1hESC) called by the AlleleDB pipeline [113]. We also created a negative training set from non-ASB variants and a randomly selected background set. Details of these sets were described in our previous study [39]. In total, we included 14,773 variants in our training set. The complementary ASB data of IMR-90 and H9 used for evaluation were downloaded from Adastra [114] at `https://adastra.autosome.org/bill-cipher/search/advanced?fdr=0.05&es=0&cl=IMR90%20(lung%20fibroblasts)` and `https://adastra.autosome.org/bill-cipher/search/advanced?fdr=0.05&es=0&cl=H9`.

4.3.2 Model architecture

TLand is a one-layer stacked architecture that consists of two parts: three base classifiers and one meta-classifier (Fig 1b and Fig 4.1). TLand takes input as 19 generic features, 40 deep learning prediction-derived features, and 5 cell-specific features or 13 organ-specific features (Table 4.1), which are directly derived from RegulomeDB queries. Features were bagged into 3 subspaces: experimental set (generic features and cell/organ-specific features), deep learning set (deep learning features and cell/organ-specific features), and cell/organ-specific set (only cell/organ-specific features). We selected lightGBM [115], random forest [116], and neural network [117] as base classifiers due to their distinct decision boundaries. Our base models were fine-tuned with Optuna [118]. We used 300 estimators in random forest models, 250 boost rounds, and a 0.049 learning rate in lightGBM models, 3 layers with 128 neurons per layer, batch size of 128, adaptive learning rate, and with the max iteration of 30 in neuron network models. Probabilities were used as the output of base classifiers and to train our meta-classifier. We calculated interaction terms of probabilities up to the degree of 2 before feeding into our meta-classifier. We customized a ridge classifier to output probabilities with hyperparameter alpha as 1.9 as our final meta-classifier. The ridge classifier was trained with 4-fold group cross-validation. We grouped the training data based on the genomic positions (i.e. variants at the same genomic positions regardless of whether cell lines were in the same group). All models were specified with balanced class weights. LightGBM was implemented with the Python package [115]. Random forest, neural network, and relevant pipelines were implemented with scikit-learn [119]. The stacked generalization algorithm was implemented with mlxtend [120].

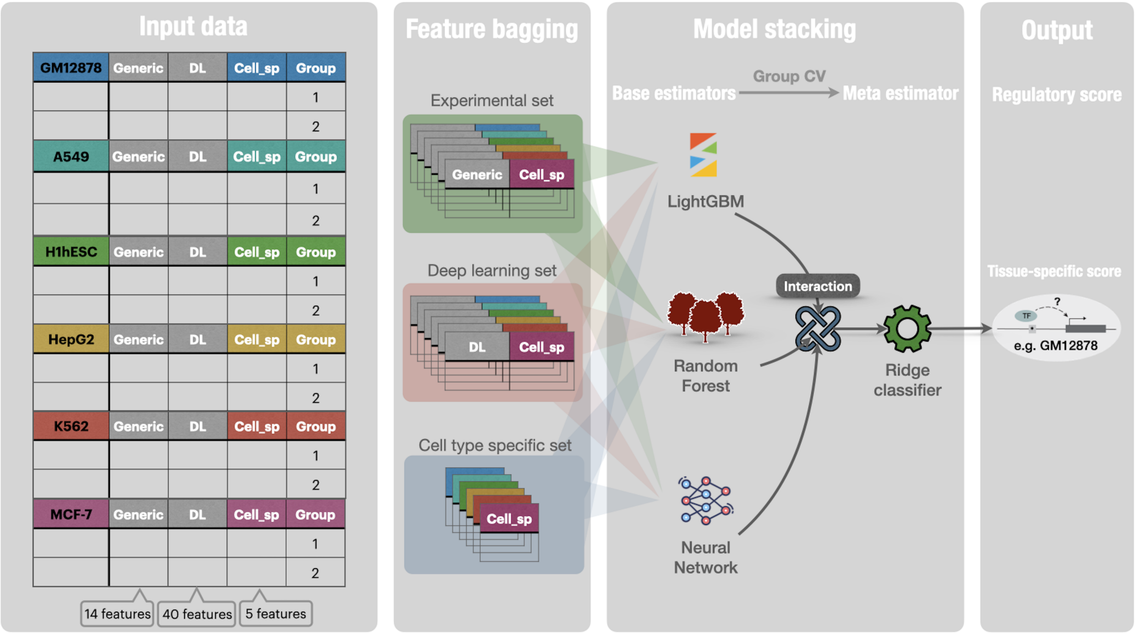
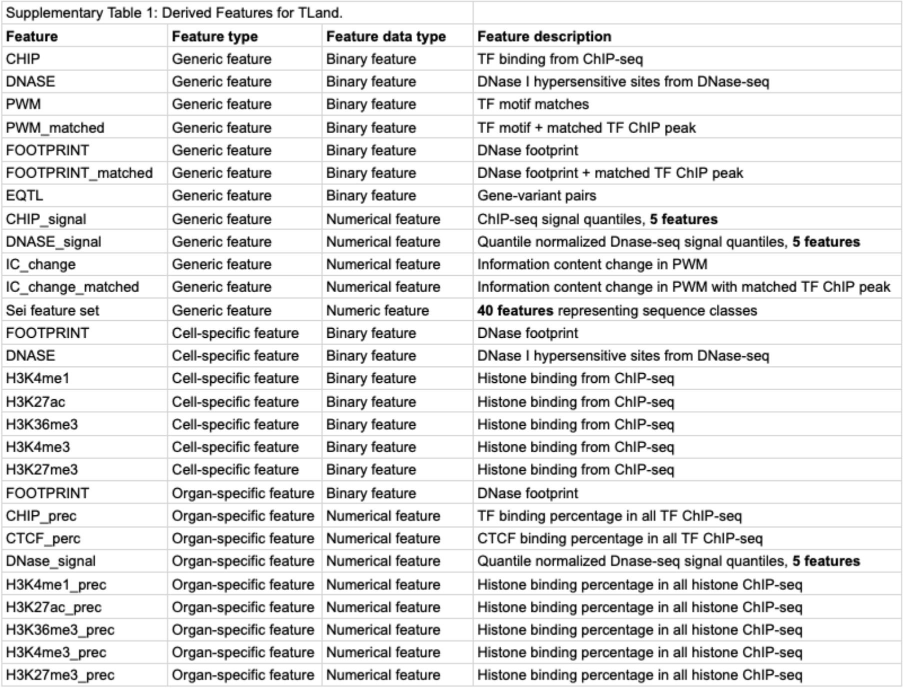


Figure 4.1: Cell-specific TLand architecture. Cell-specific TLand was trained to predict human regulatory variants in a cell-specific manner by using RegulomeDB-derived features.

4.3.3 Model training and evaluation

TLand was trained, validated, and tested on the generated ASB datasets. After concatenating across six cell lines, 88,638 (i.e. 14,773 x 6) variants were used for training and validation. The concatenation allowed us to have more data per prediction task. For the task of predicting an unseen cell line, we hold out one cell line data as a test dataset, then used the rest as a training dataset to train a TLand model. The test and train split ratio for negative data is 1/6. Similarly, we hold out one organ as a test dataset and used organ-specific labels rather than cell-specific

Table 4.1: Derived Features for TLand models.



Supplementary Table 1: Derived Features for TLand.

Feature	Feature type	Feature data type	Feature description
CHIP	Generic feature	Binary feature	TF binding from ChIP-seq
DNASE	Generic feature	Binary feature	DNase I hypersensitive sites from DNase-seq
PWM	Generic feature	Binary feature	TF motif matches
PWM_matched	Generic feature	Binary feature	TF motif + matched TF ChIP peak
FOOTPRINT	Generic feature	Binary feature	DNase footprint
FOOTPRINT_matched	Generic feature	Binary feature	DNase footprint + matched TF ChIP peak
EQTL	Generic feature	Binary feature	Gene-variant pairs
CHIP_signal	Generic feature	Numerical feature	ChIP-seq signal quantiles, **5 features**
DNASE_signal	Generic feature	Numerical feature	Quantile normalized Dnase-seq signal quantiles, **5 features**
IC_change	Generic feature	Numerical feature	Information content change in PWM
IC_change_matched	Generic feature	Numerical feature	Information content change in PWM with matched TF ChIP peak
Sei feature set	Generic feature	Numeric feature	**40 features** representing sequence classes
FOOTPRINT	Cell-specific feature	Binary feature	DNase footprint
DNASE	Cell-specific feature	Binary feature	DNase I hypersensitive sites from DNase-seq
H3K4me1	Cell-specific feature	Binary feature	Histone binding from ChIP-seq
H3K27ac	Cell-specific feature	Binary feature	Histone binding from ChIP-seq
H3K36me3	Cell-specific feature	Binary feature	Histone binding from ChIP-seq
H3K4me3	Cell-specific feature	Binary feature	Histone binding from ChIP-seq
H3K27me3	Cell-specific feature	Binary feature	Histone binding from ChIP-seq
FOOTPRINT	Organ-specific feature	Binary feature	DNase footprint
CHIP_prec	Organ-specific feature	Numerical feature	TF binding percentage in all TF ChIP-seq
CTCF_perc	Organ-specific feature	Numerical feature	CTCF binding percentage in all TF ChIP-seq
DNase_signal	Organ-specific feature	Numerical feature	Quantile normalized Dnase-seq signal quantiles, **5 features**
H3K4me1_prec	Organ-specific feature	Numerical feature	Histone binding percentage in all histone ChIP-seq
H3K27ac_prec	Organ-specific feature	Numerical feature	Histone binding percentage in all histone ChIP-seq
H3K36me3_prec	Organ-specific feature	Numerical feature	Histone binding percentage in all histone ChIP-seq
H3K4me3_prec	Organ-specific feature	Numerical feature	Histone binding percentage in all histone ChIP-seq
H3K27me3_prec	Organ-specific feature	Numerical feature	Histone binding percentage in all histone ChIP-seq

labels for splitting the data. After evaluation, the final models were trained with all 88638 variants data.

4.3.4 DNase signal quantile normalization

All 1372 DNase bigwig default files on ENCODE processed on GRCh38 assembly (up to May 2022) were obtained from `https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_title=DNase-seq&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&perturbed=true&assembly=GRCh38&files.file_type=bigWig&type=Experiment&files.analyses.status=released&files.preferred_default=true`. We have designed an efficient pipeline to quantile normalize all signals to achieve a balance between accuracy, runtime, and storage (Fig 4.2). BigWig files were first converted to BedGraph files using bigWigToBedGraph [121]. Genome-wide non-overlapping 10bp average signals were extracted from each BedGraph file using bedmap [122]. Output bed files were concatenated and then converted to parquet format as the input format of qnorm [123]. qnorm is a Python package that enables us to quantile normalize excessively large files by implementing incremental quantile normalization with multi-core support [123]. We quantile normalized all 1372 DNase signals with 10 bp bins, and with a batch size of 343 files per normalization iteration. Details of storage, runtime, and memory are shown in Fig 4.2.

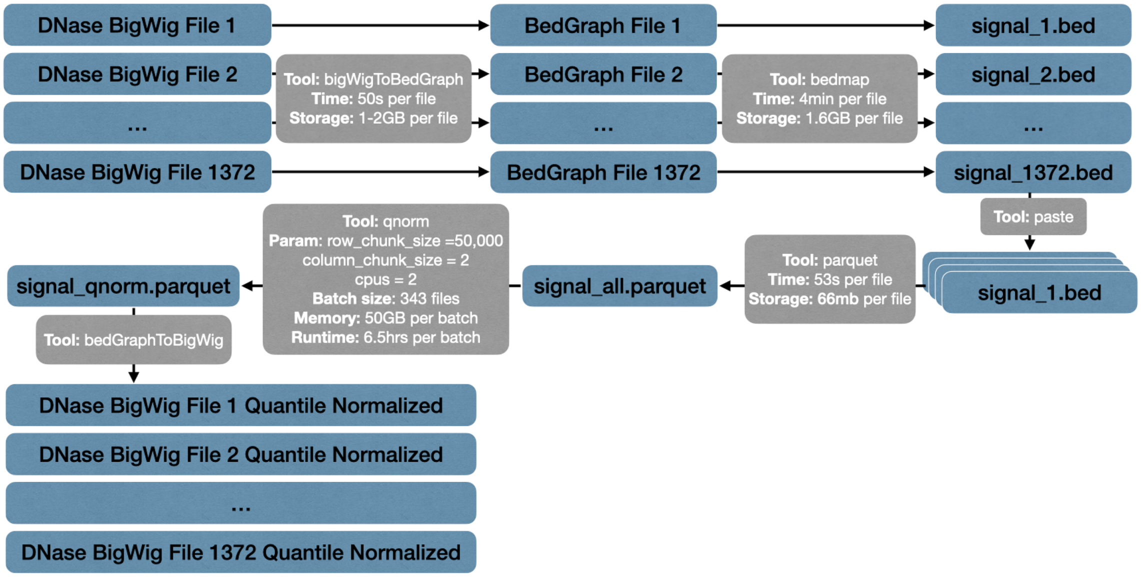


Figure 4.2: Quantile normalization pipeline. We designed a pipeline to quantile normalizes 1372 DNase BigWig files in a bin size of 10 base pairs. Important details to reproduce, including tools, parameters of tools, time, and storage it took, are all specified in the figure.

4.3.5 Benchmarking other models and organ definition

We benchmarked TLand with state-of-the-art models, including GenoNet, DeepSEA, Sei, and TURF. The pre-calculated scores from GenoNet were downloaded from `https://zenodo.org/record/3336209/files/`. The DeepSEA and Sei models were downloaded from the original paper [44, 47]. The previous best model, TURF, was re-trained to predict regulatory variants on GRCh38 by querying and inputting the GRCh38 features [39]. Final predictions were made by averaging predictions of cell lines instead of the test unseen cell line models. We averaged relevant cell line model predictions to estimate the organ level predictions. For example, we averaged TURF model predictions of A549 and IMR-90 to predict lung regulatory variants.

We used the human organ definition from ENCODE: https://www.encodeproject.org/summary/?type=Experiment&control_type!=*&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released.

4.3.6 GWAS data and LD data

We downloaded all variants associated with GWAS traits from GWAS Catalog (https://www.ebi.ac.uk/gwas/api/search/downloads/alternative) and focused on SNVs. We completed LD-expansion for each SNV, where we included SNVs from the 1000 genome project that is in strong LD (R2 threshold of 0.6). The R2 values were downloaded from (s://genomics-public-data/linkage-disequilibrium). In total, we included 1,974,549 SNVs and calculated their TLand model scores.

4.3.7 Target organs annotations for GWAS traits

Target organs for 44 GWAS traits were annotated with Open Targets (https://platform.opentargets.org/), EMBL-EBI Ontology Lookup Service (https://www.ebi.ac.uk/ols/index), and ChatGPT [124] for selected GWAS traits (See Supplementary Table 3 at https://docs.google.com/spreadsheets/d/1xYbwpp7GLHeBYip36Q-Nm0ewdgrcT-IxG-jRJACPr3M/edit?usp=share_link). Annotations were the gold standard to evaluate models for given GWAS traits and SNPs.

4.3.8 Prioritization of relevant organs for GWAS traits

We predicted on 2 million GWAS, and within the same LD block, SNPs across 51 organs. For each manually annotated GWAS trait, we selected all associated SNPs associated. For each organ, we calculated the p-value from a one-tail t-test [125] comparing the sampled organ-specific SNP scores vs. the population distribution, which was defined as the 2 million SNP score distribution for the organ. We ranked organs by their p-values, the lower the p-value, the higher the rank. We set the p-value threshold as 0.05, i.e. we filtered out organs from the ranking whose p-value was higher than 0.05. We compared the ranked (i.e. prioritized) organs with the manually annotated target organs for each trait. Accuracy was defined as the number of overlapping organs between prioritized organs and target organs, then divided by the number of target organs for a given trait. We only selected the top 4 organs for each model to calculate accuracy. We combined any two models ranking lists by taking the union of their list then ranked by p-value, and selected the top 5 organs for each combination.

4.3.9 GWAS trait enrichment

Top-scoring GWAS variants for each organ were selected; specifically, all variants with organ-specific scores >0.5 were chosen. For each organ, we traced back to the traits that those top variants are associated with, then counted the number of appearances of each trait for those variants and weighted the count with organ-specific scores. We then filtered out traits whose total count (i.e. trait count of all associated GWAS SNPs plus SNPs within 0.6 LD) was low for each organ. Organ-specific GWAS trait enrichment score was calculated as the percentage of the weighted number of appearances to the total count.

4.4 Results

4.4.1 TLand incorporates comprehensive datasets to predict regulatory variants

The continuously growing amount of genomic data portrays a more and more comprehensive and complete picture of the human regulatory variants map. RegulomeDB recently upgraded to version 2 which expanded to >650 million and >1.5 billion genomic intervals in hg19 and GRCh38, respectively [103]. The large discrepancy in the data availability between the two assemblies, for example in ChIP-seq and open chromatin data (RegulomeDB TF ChIP-seq availability in Fig 4.3a, open chromatin in Fig 4.4, histone ChIP-seq in Fig 4.5), is due to ENCODE mapped newly generated data to GRCh38 for better representation of complex variation and correction of sequencing artifacts [5]. This indicates that GRCh38 is a more resourceful assembly to derive features for training models. The features used in the new architecture to predict regulatory variants, including experimental and computational features (or evidence), were derived in GRCh38 and the majority of features were accessed and derived from RegulomeDB v2 (Table 4.1). 1372 DNase-seq BigWig files were quantile-normalized by 343 files per batch and derived as 5 quantile features for each genomic locus. The computational feature DeepSEA disease score [44] was substituted by its successor Sei model features. The Sei model simultaneously predicted 21,907 binary assay labels which were dimension-reduced to 40 features representing sequence classes [47]. Assembling deep learning model predictions and training with comprehensive features reduce variances of our final models and generalizes to new cell lines or less studied organs [126, 127].

We developed a new model architecture, TLand, to predict regulatory variants from a comprehensive set of features derived from RegulomeDB (Fig 4.3b and Fig 4.1). TLand takes input as genomic positions or dbSNP IDs, queries RegulomeDB for features, then outputs the probabilities of variants as regulatory variants in a cell-type specific (Fig 4.1) or organ-specific manner (Fig 4.3b). Training data across six common cell lines were concatenated to train an agnostic model while the specificities of input features decided the cell or organ specificities for model predictions.

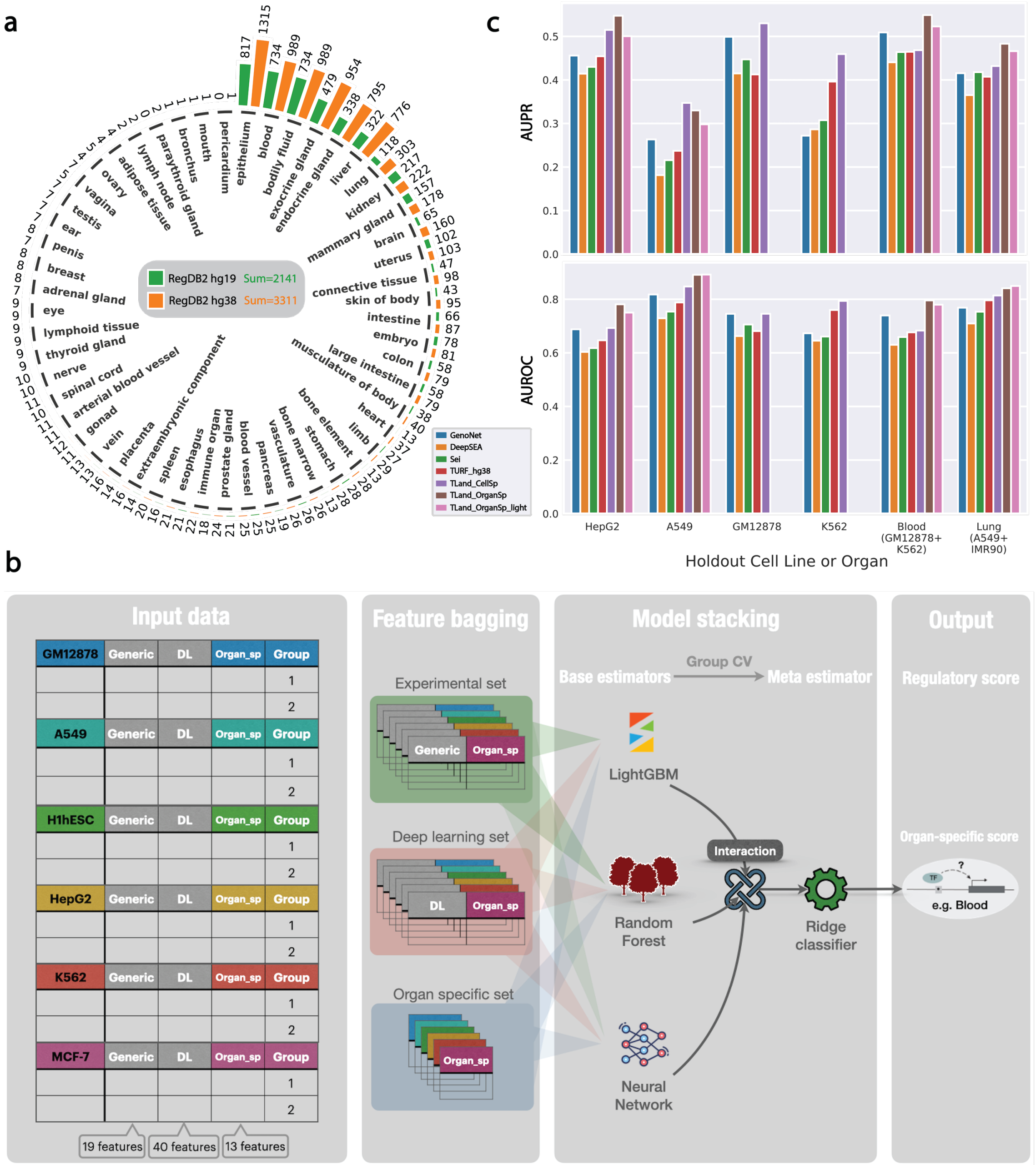


Figure 4.3: TLand improves regulatory variant predictions. (a) TF ChIP-seq data availability across organs on RegulomeDB v2. Green bar plots represent counts on GRCh38. Orange bar plots represent counts on hg19. The total number of counts for each assembly is in the middle of the gray box. Notice that the summation is not simply adding all numbers together due to some cell lines having multiple corresponding organs. (b) Organ-specific TLand architecture. Organ-specific TLand was trained to predict human regulatory variants in an organ-specific manner by using RegulomeDB-derived features. (c) Benchmarking TLand performance by AUROC and AUPR. X-axis is holdout cell lines or organs. Y-axis is AUPR on the top panel and AUROC on the bottom panel.

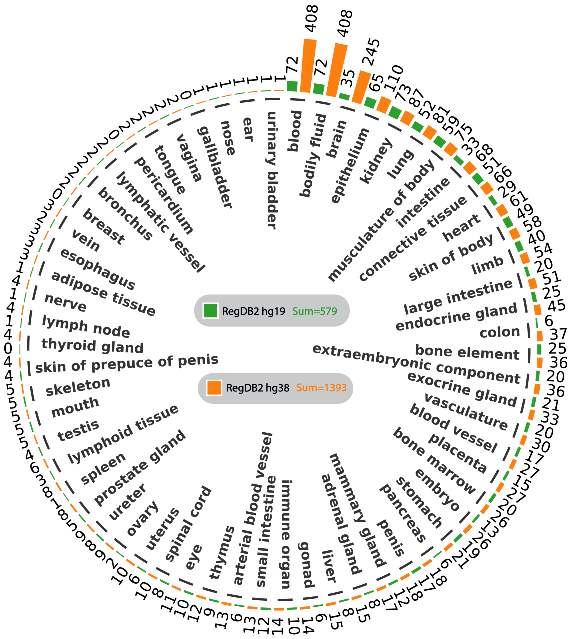


Figure 4.4: DNase-seq data availability across organs on RegulomeDB v2. Green bar plots represent counts on GRCh38. Orange bar plots represent counts on hg19. The total number of counts for each assembly is in the middle of the gray box. Notice that the summation is not simply adding all numbers together due to some cell lines having multiple corresponding organs.

Features were bagged into three biologically meaningful subspaces before each set was fed into an individual base classifier. Subspaces included experimental features set, computational features set, and cell/organ-specific features set for regulatory variants. We adopted the stacked generalization algorithm to stack the output of individual classifiers and use a meta-classifier to compute the final prediction [55]. Stacking allows for the utilization of the strength of each individual classifier by using their output as input of a final meta-classifier. Cross-validation is required to prevent information leaks during the training of the meta-classifier. We made several modifications to the algorithm. We used group cross-validation to make base classifiers to learn the regulatory function as the conditional probability between generic features and cell/organ-specific features. Interaction terms were calculated before feeding them into the meta-classifier. We used probability rather than binary

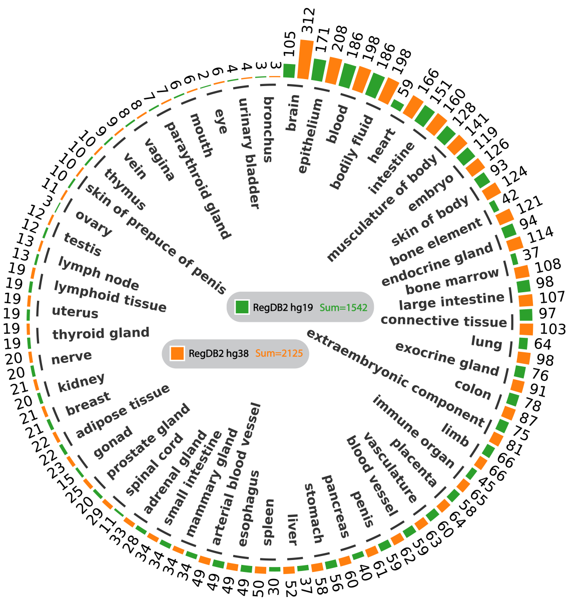


Figure 4.5: Histone ChIP-seq data availability across organs on RegulomeDB v2. Histone ChIP-seq data includes five histone marks: H3K4me1, H3K27ac, H3K36me3, H3K4me3, and H3K27me3. Green bar plots represent counts on GRCh38. Orange bar plots represent counts on hg19. The total number of counts for each assembly is in the middle of the gray box. Notice that the summation is not simply adding all numbers together due to some cell lines having multiple corresponding organs.

prediction in the first layer of base classifiers to train the meta-classifier to obtain higher accuracy [128].

4.4.2 TLand improves regulatory variant predictions

Cell-specific TLand models substantially outperformed state-of-the-art models for predicting unseen cell line regulatory variants (Fig 4.3c and Fig 4.6). On average, across holdout cell lines, cell-specific TLand models outperformed the previous best model TURF, with the area under the precision-recall curve (AUPR) and the area under the receiver operating characteristic (AUROC) increasing from 0.389 to 0.471, and 0.729 to 0.776, respectively. Although GenoNet outperformed TURF in three cell lines, the higher performance was due to information leakage (i.e. we used the unseen cell line model of GenoNet to predict and assess the upper limit of the prediction task). Cell-specific TLand models still had superior performance compared to GenoNet, improving average AUPR by 0.09 and AUROC by 0.04. The high performance of cell-specific TLand can be explained in two ways. First, we derived comprehensive feature sets for predicting regulatory variants from RegulomeDB, categorizing them into experimental features set and deep learning features set which are not correlated, but complementary to each other to make more accurate predictions (Fig 4.7). When only deep learning features were used, such as in DeepSEA and Sei models (average AUPR 0.336 and 0.364), both largely underperformed when compared to cell-specific TLand, which combines experimental features with deep learning features (average AUPR 0.471). The other explanation is due to TLand's architecture, which models different biological domains separately and then calculates their conditional probabilities. The meta-classifier was trained to find the best combinations of the probabilities and subtracted the common redundant information (i.e. the negative coefficients in Fig 4.8) to de-noise and thus improve the prediction. The advantage of this architecture was evidenced by comparing TLand and TURF in the original hg19-derived feature context. TLand improved average AUPR by 0.10 and average AUROC by 0.04 compared to TURF (Fig 4.9). The flexible architecture of TLand allows for the substitution of cell-specific features with organ-specific features, which can be retrained with organ-specific labels to develop organ-specific TLand models.

4.4.3 Organ-specific TLand models address data availability bias

Organ-specific TLand models have the ability to predict regulatory variants specific to organs, even those from less-studied cell lines or tissues within the organ. These models performed better than their cell-specific counterparts in two out of four cell-type specific tasks, specifically in HepG2 (AUPR 0.547 vs. 0.514, AUROC 0.781 vs. 0.692) and MCF-7 (AUPR 0.512 vs. 0.512, AUROC 0.879 vs. 0.840).

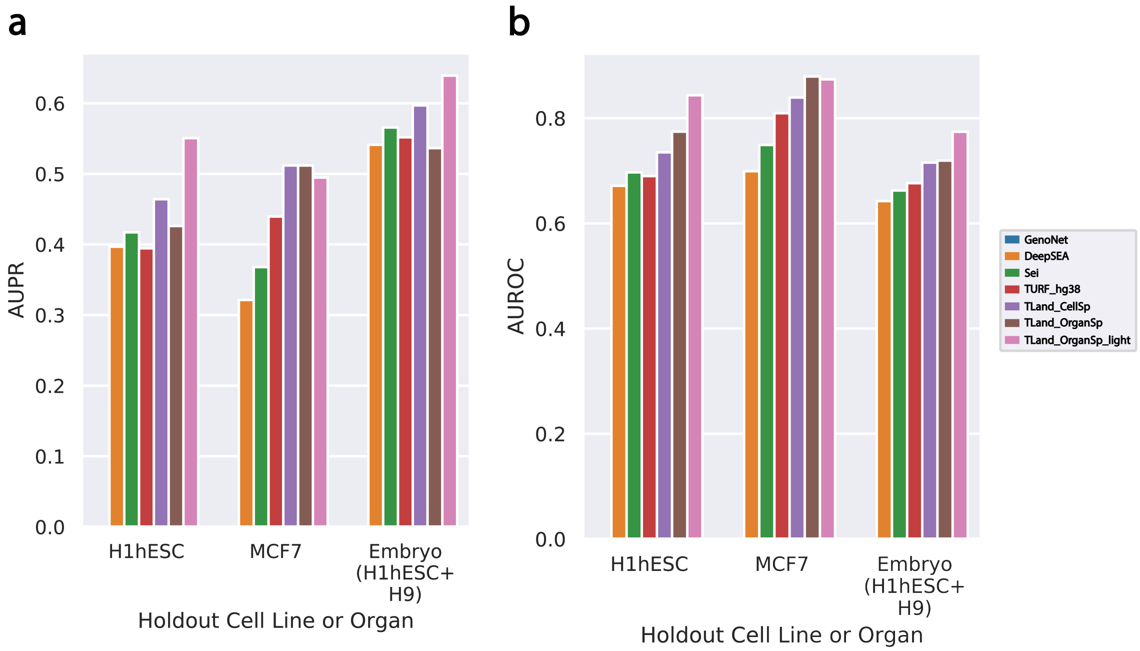


Figure 4.6: Benchmarking models on H1, MCF7, and embryo. X-axis is holdout cell lines or organs. Y-axis is AUPR on the left panel and AUROC on the right panel.

However, GM12878 and K562 cell lines lacked organ-specific models due to conflicting training labels and were thus assessed within the blood organ. The superior performance of organ-specific models, even in hold-out cell line tasks, can be attributed to the fact that cell-type specific regulatory variants in HepG2 and MCF-7 are well-represented by corresponding organ-specific regulatory variants within the RegulomeDB database (see Fig 4.10). In the A549 cell line, which is the least represented lung organ, the organ-specific TLand model underperformed the cell-specific TLand model holdout on A549 (AUPR 0.329 vs. 0.347) as the organ model predicted regulatory variants in the lung other than in A549 cell line. To better evaluate model performance holdout in the lung, the ASB dataset from the second-most representative cell line in the lung, IMR-90, was added to complement the holdout dataset. The inclusion of this dataset led to TLand outperforming the cell-specific model when holding out on the lung with IMR-90 datasets (AUPR 0.483 vs. 0.432, AUROC 0.841

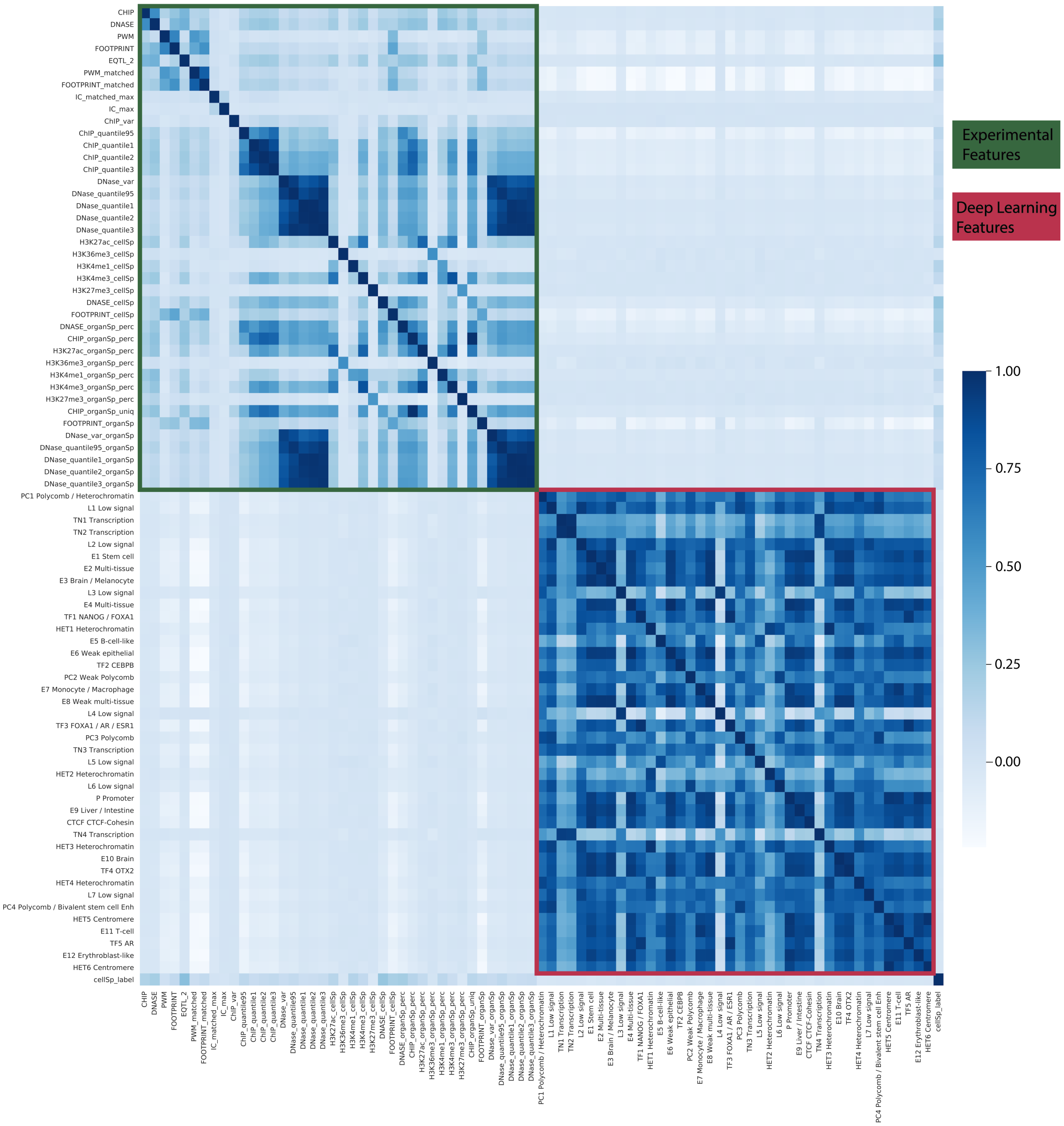


Figure 4.7: Correlation between deep learning features and experimental features in GM12878 cell line. Both axes are features derived from RegulomeDB for the GM12878 ASB dataset. The heatmap was calculated as the correlation between those features, ranging from [-1, 1]. The green box represents the features belonging to the experimental feature set. The red box represents the features belonging to the deep learning feature set.

vs. 0.814), indicating the organ-specific TLand model can predict a comprehensive set of organ-specific regulatory variants.

However, adding the second most representative cell line in the embryo H9, did not improve the TLand organ-specific model when compared to the cell-specific model

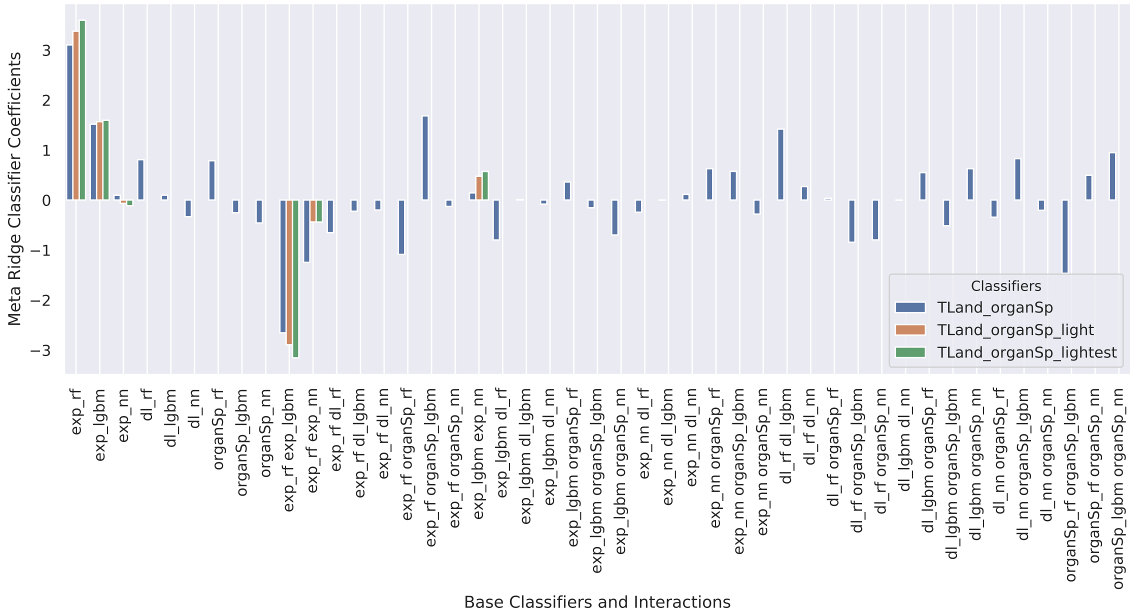


Figure 4.8: Ridge classifier coefficients. The X-axis is base classifiers and their interaction term from the final TLand models trained on all data. Y-axis is the coefficient of the meta-ridge classifier. Blue represents TLand organ-specific model. Orange represents TLand organ-specific light model which did not have deep learning features (i.e. no dl base classifiers nor organSp base classifiers). Green represents TLand organ-specific lightest model which removed organ-specific ChIP-seq features from the light model.

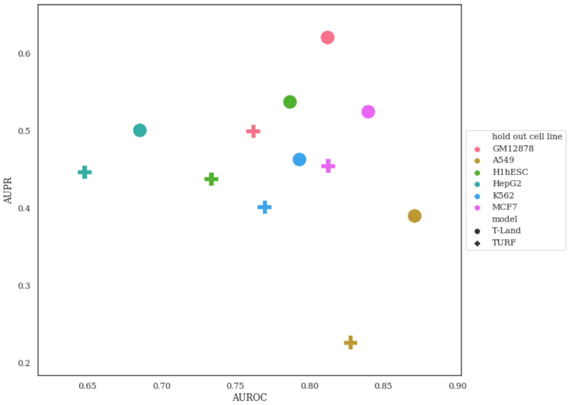


Figure 4.9: Benchmarking TLand and TURF on hg19. X-axis is AUROC, and Y-axis is AUPR. Colors represent holdout cell lines. Shapes represent models. Specifically, the circle represents the cell-specific TLand model, and the plus sign represents the TURF model.

when holding out on the embryo organ (AUPR 0.536 vs. 0.597). Deep learning model features, such as DeepSEA disease impact score [44] and Sei sequence classes [47], were more representative of the cell lines or organs with more data availability.

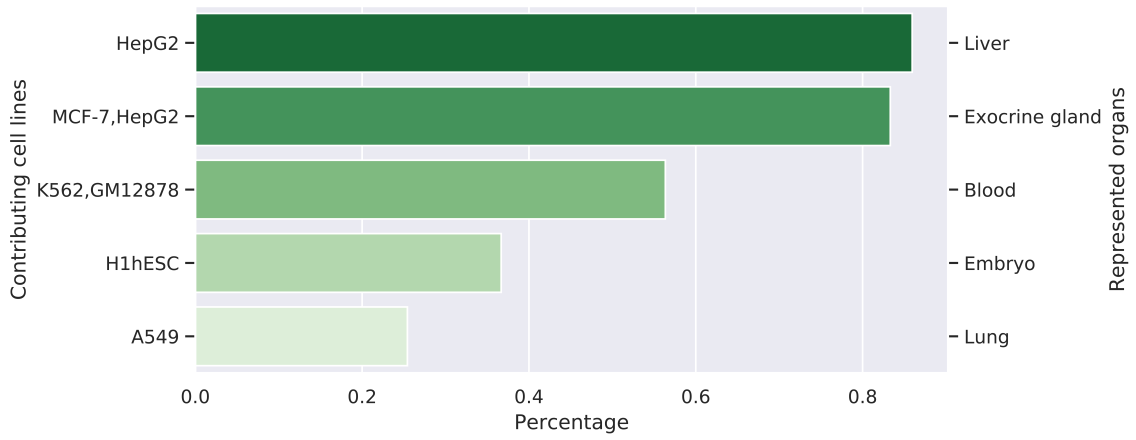


Figure 4.10: Organ representability by cell lines on RegulomeDB v2. X-axis is the representability percentage of organs, which indicates how well each organ is represented in the RegulomeDB by the contributing cell lines. The higher the representability, the darker the green color. Y-axis on the left is contributing cell lines. Y-axis on the right is the represented organs.

This is because the scores were dimensionally reduced from 919 and 21,907 prediction tasks of experimental assays for DeepSEA and Sei, respectively (including TF and histone ChIP-seq and open chromatin assays). The more representative cell lines or organs regarding data availability dominated the dimension-reduced results. Thus, we removed deep learning features and developed the organ-specific TLand light model (Fig 4.11) to predict regulatory variants accounting for low data availability, such as in the embryo organ. We defined organs with high data availability if their number of TF ChIP-seq assays were more than 100, and with low data availability if below 100. We observed that the TLand (full) model consistently outperformed the TLand light model in organs with high data availability (Fig 4.3a and c; Fig 4.6), while the light model surpassed the full model in organs with low data availability. For example, the organ-specific TLand light model was the best model when holding out on the embryo organ (AUPR 0.639, AUROC 0.774). Those findings indicate that TLand light models are suitable for predicting regulatory variants for organs with low data availability, and TLand (full) models are suitable for organs with high data availability. After evaluation, we proceeded with our analysis by training the TLand model and the TLand light model on all data. We trained an additional TLand model, TLand lightest, where the organ-specific ChIP-seq features were removed to further reduce bias towards over-represented organs (benchmarking results shown in Supplementary Table 2 at `https://docs.google.com/spreadsheets/d/1xYbwpp7GLHeBYip36Q-NmOewdgrcT-IxG-jRJACPr3M/edit?usp=share_link`).

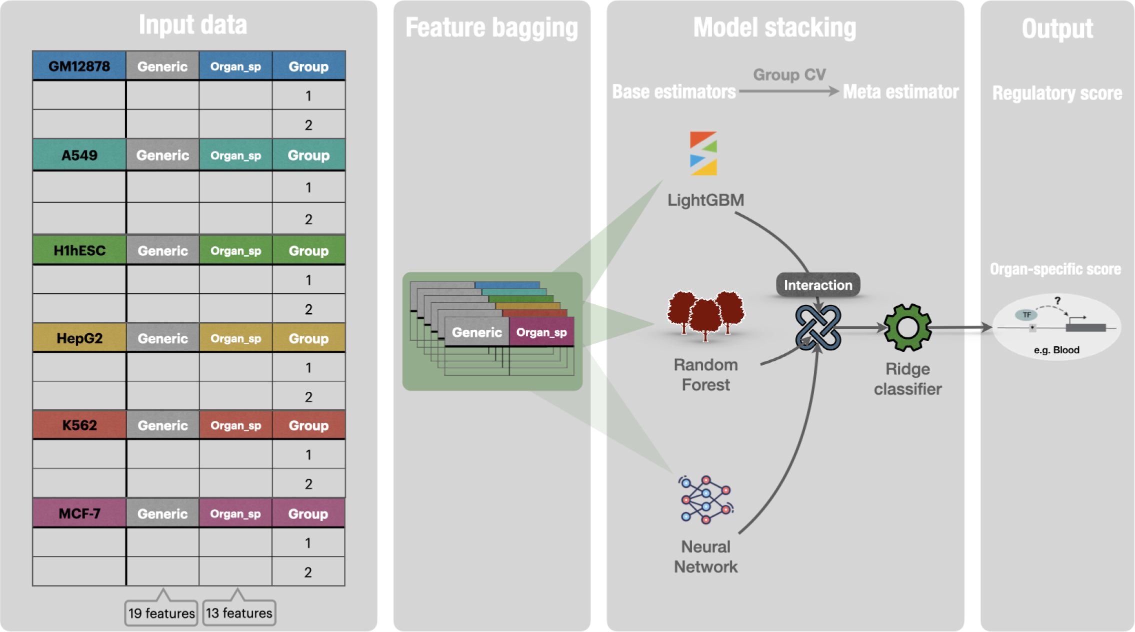


Figure 4.11: Organ-specific TLand light architecture. Organ-specific TLand light was trained to predict human regulatory variants in an organ-specific manner by only using RegulomeDB-derived experimental features.

4.4.4 TLand prioritizes relevant organs for GWAS traits

To systematically evaluate organ-specific TLand models, we predicted on around 2 million GWAS SNPs including SNPs within the same LD blocks across 51 organs defined by ENCODE [5]. We found that TLand predictions were more highly correlated with TLand light predictions than TLand lightest (e.g. in the heart organ in Fig

4.12). In addition, TLand models predicted clustered organs of the same or closely related system. For example, TLand (Fig 4.13a) and TLand light (Fig 4.14) clustered the brain and spinal cord from the central nervous system, and the eye and ear from the sensory system together. TLand lightest model, however, was able to cluster the blood vessel, the arterial blood vessel, and the vascular together as part of the circulatory system (Fig 4.15), which was missed by the other two models. This indicates that various types of TLand models attended to different organs varying on data availability. We manually curated a list of GWAS traits with relevant organs. TLand, TLand light, and TLand lightest models prioritized the relevant organs for 44 GWAS traits (See Supplementary Table 3 at `https://docs.google.com/spreadsheets/d/1xYbwpp7GLHeBYip36Q-NmOewdgrcT-IxG-jRJACPr3M/edit?usp=share_link`) with an average accuracy of 0.311, 0.340, and 0.466, respectively (Fig 4.13b. See the definition of accuracy in Methods). By integrating prioritized organs from two distinct models, TLand and TLand lightest, the average accuracy was increased to 0.482. The overall low accuracy was partially due to the uncertainty of a low score, whether it was caused by low data availability or if it was a true negative (i.e. not a functional regulatory variant in the organ), especially for the organs with low data availability.

Focusing on top-scoring variants for each organ, however, allowed us to attend to the precision of TLand models, which is a more appropriate metric to evaluate the low data availability organ models. We found that top-scoring variants by organ-specific models were enriched with associated GWAS traits that were relevant to the organs (Methods, Fig 4.13c and all results in Supplementary Table 4 at `https://docs.google.com/spreadsheets/d/1xYbwpp7GLHeBYip36Q-NmOewdgrcT-IxG-jRJACPr3M/edit?usp=share_link`). For example, atrial fibrillation and resting heart

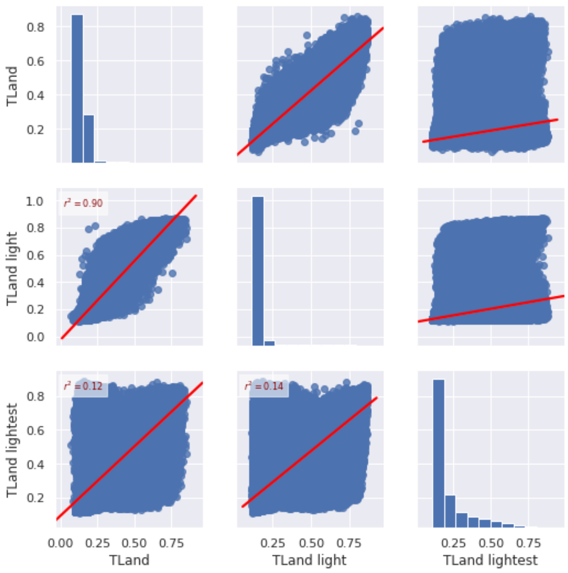


Figure 4.12: Pairplot of TLand model prediction scores in the heart. X-axis and Y-axis are TLand, TLand light, and TLand lightest model prediction scores in order.

rate were identified as two of the three most enriched traits among the total of 2804 GWAS traits by the TLand lightest model (note that not all traits were shown in Fig 4.13c). TLand lung models were not able to prioritize the lung organ for lung-relevant traits given corresponding GWAS SNPs previously (See Supplementary Table 3 at `https://docs.google.com/spreadsheets/d/1xYbwpp7GLHeBYip36Q-Nm0ewdgrcT-IxG-jRJACPr3M/edit?usp=share_link`). However, we found that traits such as physical activity measurement and peak expiratory flow were enriched by top-scoring variants in the TLand model. Traits that were associated with multiple organs were able to be pinpointed by multiple organ models. Ankylosing spondylitis, which is a type of inflammatory arthritis, affects the spine and large joints [129]. Both the bone element and the immune organ TLand lightest model pinpointed ankylosing spondylitis as one of the most enriched GWAS traits in their top-scoring variants (Fig 4.13c), indicating organ-specific TLand models can accurately prioritize relevant organs given regulatory variants.

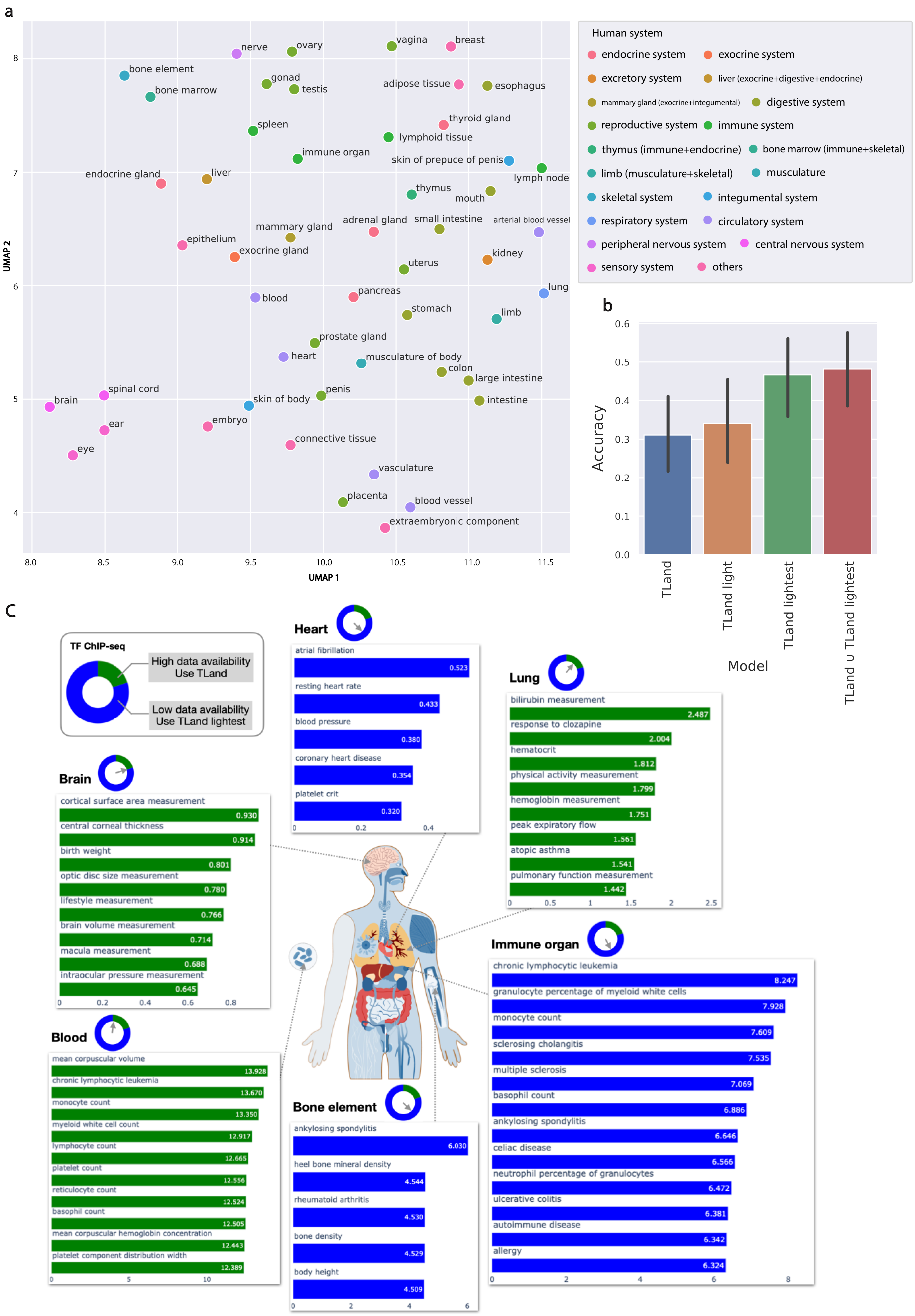


Figure 4.13: TLand prioritizes relevant organs for GWAS traits. (a) UMAP projections of TLand predictions on GWAS and LD SNPs. They are colored according to human systems. (b) Performance of TLand models prioritizing organs. X-axis are TLand models. Y-axis is accuracy (definition see Methods). (c) Prioritization of GWAS traits given top-scored variants by TLand and TLand lightest models. TF ChIP-seq data availability plot is re-plotted in the top left where green represents high data availability and TLand was the model used, and blue represents low data availability and TLand lightest was the model used. Each circle next to the organs indicates the data availability of each organ. The colors of enrichment bars correspond to the model used.

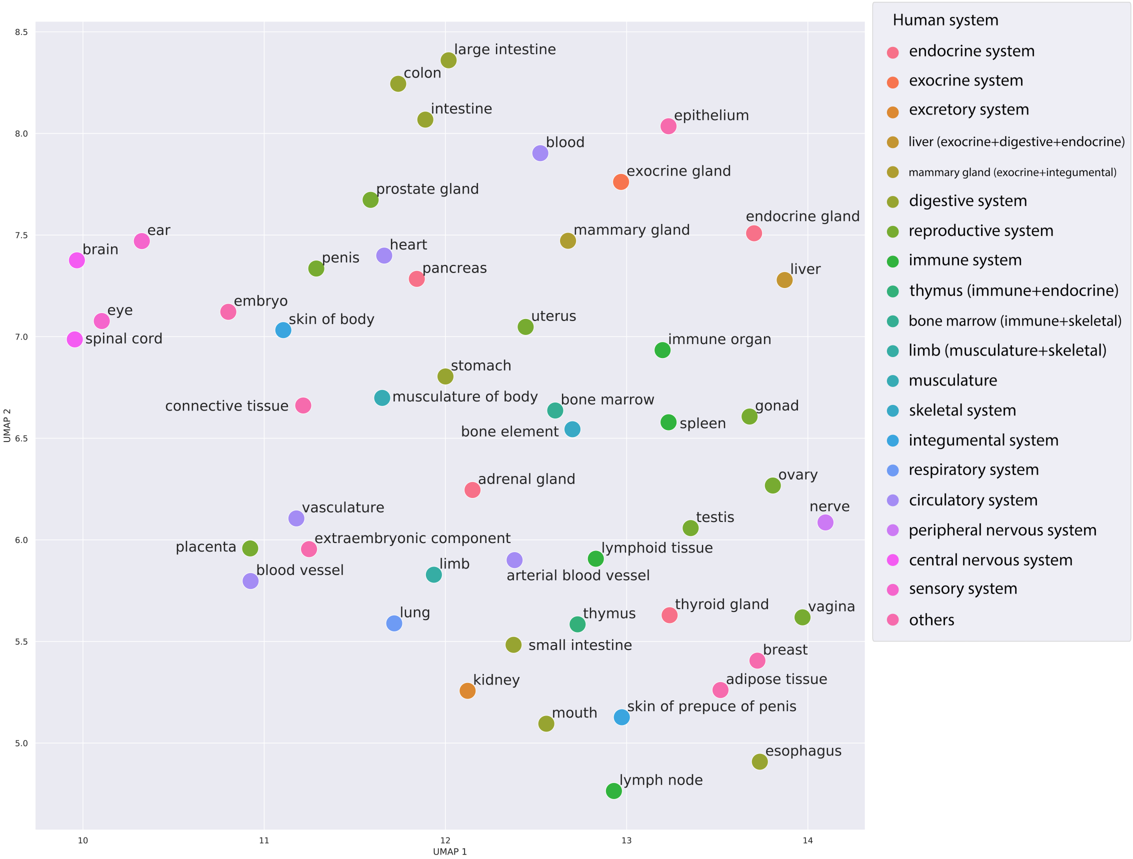


Figure 4.14: UMAP projections of TLand light predictions on GWAS and LD SNPs. They are colored according to human systems.

4.5 Discussion

The identification and characterization of non-coding regulatory variants in the human genome is still a major challenge in the field of genomics. With the recent advancement of RegulomeDB, we have developed a novel model architecture, TLand, which utilizes the RegulomeDB-derived features to infer regulatory variants at multiple levels, either at the cell-specific or organ-specific level. By incorporating comprehensive datasets on GRCh38 with stacked generalization, including experimental and computational features, TLand models were able to consistently outperform state-of-the-art models in holdout benchmarking, demonstrating TLand models' ability to

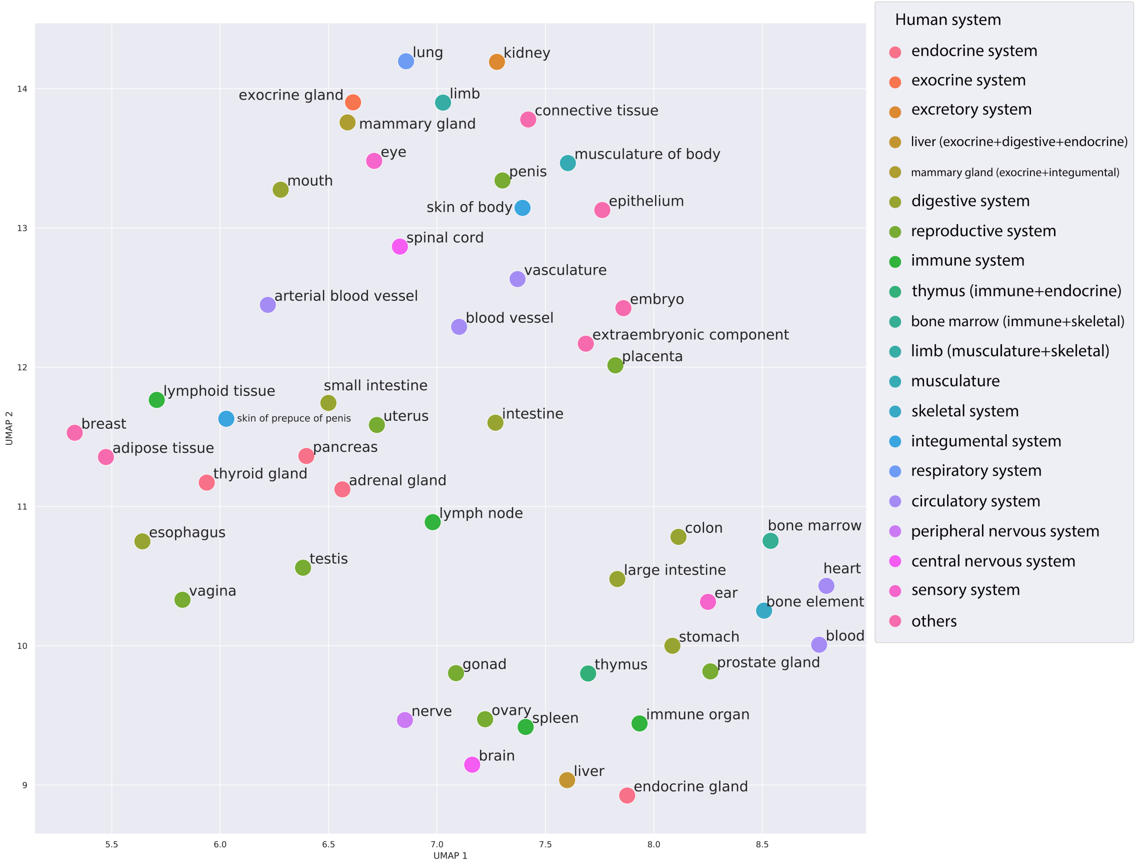


Figure 4.15: UMAP projections of TLand lightest predictions on GWAS and LD SNPs. They are colored according to human systems.

generalize to new cell lines or organs. We accounted for the data availability issue of various organs, a common issue observed in models such as DeepSEA disease impact score, by training three types of TLand models to attend to organs with high and low data availability. Combining TLand models enabled us to prioritize the relevant organs for GWAS traits accurately.

There are several ways to further improve TLand models. Machine learning models' success depends on the training data. One limitation of this study is the limited number of allele-specific binding (ASB) sites in our training dataset. We generated our training datasets by using personal genomes of only six common cell lines due to the limited data availability of other cell lines, which could result in less confident ASB calling for other cell lines. In addition, there is a large disagreement between ASB calling methods, which leads to an even smaller set of confident training datasets. Recently, a new database, Adastra, has been published [114]. It hosts 652,595 ASBs passing 5% FDR across 647 cell lines and 1,043 TFs. However, being able to call such a comprehensive set of ASBs, the authors had to use statistical distributions rather than the more accurate personal genomes to call ASBs. Whether we could use the dataset to improve TLand depends on the further evaluation of the quality of Adastra ASB calling. Moreover, we aim to continuously incorporate new datasets into RegulomeDB which could be derived as new features for TLand models, such as gkm-SVM model predictions [41] and Hi-C contact maps [112]. Due to the flexible modularity design of TLand architectures, we are able to group features into biological meaningful sets or create a new feature set, monitor and evaluate them modularly before making the decision about whether we would include them in the model.

Despite the aforementioned limitation, TLand presents a valuable contribution to the field of non-coding variant analysis by incorporating comprehensive datasets to predict regulatory variants at a cell-specific or organ-specific level. To foster downstream applications, we have made the pre-trained TLand models available. In addition, we have made openly available pre-calculated TLand scores for the union of open chromatin regions. TLand along with RegulomeDB further advance our ability to identify human regulatory variants genome-wide, and the model's flexibility allows for further integration of data and features as they become available.

4.6 Publication

The work described in this chapter is being prepared for publication. I will be the first author of the paper.

CHAPTER V

Identifying Patterns in Genomic Sequences with Deep Learning Models

5.1 Abstract

Interpreting predictive machine learning models to derive biological knowledge is the ultimate goal of developing models in the era of genomic data exploding. Recently, sequence-based deep learning models have greatly outperformed other machine learning models such as SVM in genome-wide prediction tasks. However, deep learning models, which are black-box models, are challenging to interpret their predictions. Here we represented an end-to-end computational pipeline, Explain-seq, to automate the process of developing and interpreting deep learning models in the context of genomics. Explain-seq takes input as genomic sequences and outputs predictive motifs derived from the model trained on sequences. We demonstrated Explain-seq with a public STARR-seq dataset of the A549 human lung cancer cell line released by ENCODE. We found our deep learning model outperformed the gkm-SVM model in predicting A549 enhancer activities. By interpreting our well-performed model, we identified 47 TF motifs matched with known TF PWMs, including ZEB1, SP1, YY1, and INSM1. They are associated with epithelial-mesenchymal transition and lung cancer proliferation and metagenesis. In addition, some motifs were not matched in the JASPAR database and may be considered *de novo* enhancer

motifs in the A549 cell line.

5.2 Introduction

Decoding regulatory functions that are encoded in genomic sequences is a major challenge in understanding how genomic variations are associated with phenotypic diseases and traits. High-throughput sequencing methods have been developed to screen for regulatory regions genome-wide. DNase I hypersensitive site sequencing (DNase-seq) [130] is designed to detect genome-wide chromatin accessibility. Transcription factor (TF) binding and histone modifications are measured using Chromatin Immuno-Precipitation sequencing (ChIP-seq) [106, 16]. Enhancers are regulatory DNA sequences that recruit TFs to up-regulate target gene expression in a cell-type-specific manner, which governs physiology and development in humans [131]. STARR-seq is a massively parallel reporter assay to identify potential enhancers and provide a direct functional or quantitative readout of enhancer activity genome-wide [131]. The genome-wide quantitative enhancer map enables interrogating enhancers in higher resolution than binary peak regions such as peaks from DNase-seq.

Deep learning techniques have made substantial progress in modeling genomic sequences to predict epigenetic marks such as DNA accessibility, TF, and histone marks across a range of cell types. Particularly, convolutional neural networks (CNNs), have been shown to accurately predict epigenomic features with DNA sequences [132, 133, 44]. For example, DeepSEA, given any 1000 bp DNA sequence, can accurately predict 919 binary labels for the sequence, representing open chromatin regions, TF binding sites, and histone mark regions altogether in a multi-label CNN model [44]. The model significantly outperformed the previous state-of-art gkm-

SVM, which leveraged a gapped-kmer-SVM classifier to predict functional sequence elements in regulatory DNA [134]. One potential explanation for why CNN is a better fit for genomic sequence learning tasks is that filters learned during training are analogous to position weight matrices (PWMs) of motifs, which are often conserved and positional invariant.

However, interpreting machine learning models, especially black-box deep learning models, to derive the biological knowledge learned by models remains elusive. Specifically, what sub-sequences make contributions to certain predictions, and how we can summarize those sub-sequences into human-readable motifs? Additionally, due to the limited quantities of conducted STARR-seq experiments, the relationship between enhancer sequences and activities across different cell lines is still poorly understood. There lacks a human enhancer sequence-to-activity model that learns the cis-regulatory grammar in a cell-type specific manner, which can accurately predict enhancer activities.

To address these questions, we presented a novel end-to-end pipeline, Explain-seq (Fig 5.1A), to automate the process of developing and interpreting deep learning models. ENCODE has recently released 6 STARR-seq datasets for common human cell lines [5]. These new datasets provide an opportunity to examine and compare enhancers in a cell-type-specific manner. Here we demonstrated Explain-seq by analyzing a new STARR-seq dataset in the A549 lung cancer cell line. The pipeline started with training a CNN with a regression layer at the end of the network to predict cell-line-specific enhancer activity and ended with outputting derived predictive motifs. Our trained regression model outperformed the gkm-SVM model in terms of the Pearson correlation coefficient. Also, derived motifs from Explain-seq were matched to other known TF motifs in the JASPAR database including ZEB1, YY1, SP1, and INSM1, which are associated with lung cancer [66]. In addition, there were derived motifs not similar to any known motifs, for example, one was enriched with short repeats ATGAAA, which may be considered as *de novo* motifs. The *de novo* motifs discovered by Explain-seq may allow us to design the synthetic enhancers with desired activity in a cell-line-specific manner.

5.3 Methods

5.3.1 Data access and preparation

ENCODE phase 4 released 6 STARR-seq datasets on June 11, 2020, which include common human cell lines MCF-7, HCT116, A549, SH-SY5Y, HepG2, and K562 [5]. We downloaded their peak regions in .bed format and signal values in .bigwig format.

To prepare input data for the deep learning model, we first selected all regions at the summit of each STARR-seq peak and binned them into 499-bp windows. We limited the input sequence length to 499bp since the size selection in STARR-seq experiments before cloning to plasmids is 500 bp [135]. We included 3 adjacent regions on either side of each selected summit region by sliding a 100 bp overlapping window. We randomly selected 500,000 regions of size 499 bp on the hg38 human genome excluding ENCODE blacklist regions as negative sets [136]. The signal for each 499-bp window was calculated by the averaging signal value of the whole corresponding region. For the regression model, the average signal values were directly used for training. For the multi-label classification model, we further generated 10 labels using the average value. In total, we have selected 1,027,953 peaks as our input data. We split our input data by chromosome for training, validation, and testing. Specifically, we used chr7 for validation, and chr8 and chr9 were held out for testing.

5.3.2 Overview of Explain-seq pipeline

Explain-seq pipeline is an end-to-end analytical pipeline to discover potential known and *de novo* motifs given genomic sequences (Fig 5.1A). It takes input as genomic region coordinates with labels for classification tasks (in .bed format) or one-hot encoded sequences with continuous value for regression task (in .mat format) [137]. In addition, it requires a user-defined deep learning model in PyTorch [138]. Optionally, weights from the pre-trained model can be transferred to the new model for transfer learning. After the training and validation loss function converges, the model and input sequences are piped into DeepLIFT to compute contribution scores with respect to enhancer activities in a single-nucleotide resolution [139]. TF-Modisco is then used to cluster and summarize motifs with the weighted input sequences weighted by importance scores [140]. Potential motifs are compared to known motifs in databases through STAMP [141]. Non-matched motifs are considered *de novo* predictive motifs in a specific cell line.

5.3.3 Model architecture, training scheme, and transfer learning

We designed our model as a CNN which takes input as one-hot encoded 499-bp long DNA sequences (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) to predict enhancer activities (Fig 5.1B). The architecture is inspired by the Beluga model, which has double convolutional layers than DeepSEA, to enable later transfer learning [142]. Specifically, our model includes 6 convolutional layers with an equal kernel size of 8 and has 320, 320, 480, 480, 640, and 640 filters for each layer, respectively. ReLU is used as the activation function in each convolutional layer. Every two convolution layers are followed by a dropout layer with a probability of 0.2 and a max pooling layer with a pooling size of 4. Those 6 convolution layers are followed by a fully connected layer with 2003 neurons. We provided two design choices for the final layer. The first option for multi-label classification is 10 nodes where each node represents a label in a specific cell line. For example, the first node representing the probability of the input sequence (or region) has a signal value in [0, 1), and the last node represents in [9, 10). Here, we used binary cross entropy as the loss function. The second option for regression setting is one node for continuous signal value prediction. Mean squared error (MSE) was used as a loss function in regression. The reason for developing and comparing the two designs is that the data might be noisy, and it may not have enough statistical power for us to learn a regression model. Binning the continuous values and learning a multi-label classification model may help denoise the data and learn a more general and representative model.

We implemented our model in PyTorch [138]. Considering this is a genome-wide learning task with large data, transfer learning is useful to speed up the training and also to improve accuracy [132, 45]. We transferred all the weights of convolutional layers from the pre-trained Beluga model as initial weights. The weights from the fully connected layer were excluded since the original Beluga model has an input size of 2,000 bp. After initialization, we fine-tuned the weights by re-training the model. To speed up the development, we used the Selene framework to facilitate the training process [137]. We used Adam optimizer with a learning rate of 1e-5 [143], a batch size of 128, and used an Nvidia Titan V with 12 GB memory GPU to develop our model.

5.3.4 Baseline comparisons

We compared our CNN-based regression model with another machine learning method gkm-SVM [134]. gkm-SVM trains gapped-kmer SVM classifiers for DNA sequences to detect functional sequence elements in regulatory DNA. We used the Pearson correlation coefficient (PCC) as a metric to compare the prediction accuracy with the true signals. In addition, we calculated the correlation between the actual signals in one biological replicate with the actual signals in another biological replicate to serve as the maximum prediction accuracy threshold for this task.

5.3.5 Derive nucleotide importance scores

DeepLIFT is an algorithm to calculate feature importance scores for neural networks by propagating activation differences [139]. We used DeepLIFT to derive importance scores for all input with signals larger than 3 with respect to the cell-line-specific activity. The advantages of DeepLIFT compared to other interpretation methods are: 1) the RevealCancel rule of DeepLIFT allows it to properly handle saturation cases while integrated-based methods may give misleading results [144]. 2) DeepLIFT is a good and faster approximation of the SHapley Additive exPlanation (SHAP) value [139]. 3) Di-nucleotide frequency shuffling mimics true genomic background to increase importance scores signal-to-noise ratio. As background, we shuffled 100 times for each input sequence while maintaining di-nucleotide frequency. The output of DeepLIFT is importance scores in nucleotide resolution, and hypothetical importance scores are similar to mutagenesis indicating what importance scores would be placed on a different base in the sequence.

5.3.6 Clustering and summarizing sub-sequences into motifs

TF-Modisco is a clustering-based algorithm to consolidate motifs from sequences with importance scores [140]. The algorithm started with finding potential motifs, named seqlets, through MEME [145]. TF-Modisco implemented a correlation alternative, continuous Jaccard similarity, to better calculate the similarity between seqlets than cross-correlation, and developed the density-adaptive distance to improve clustering on weak motifs when their distances are generally larger [140]. We specified the final motif size as 15 bp with 5 bp flanking on each side. The target seqlets false discovery rate was set to 0.15 and the max seqlets per cluster was set to 20,000. We sampled 4,900 sequences for our null Laplace null distribution. The output of TF-Modisco contains potential motifs with PWMs, contribution weight matrices with importance scores, and hypothetical score matrices.

5.3.7 Annotation with known motifs

STAMP was leveraged to annotate motifs from TF-Modisco within a known motif database, JASPAR [66, 141]. We trimmed off the motif edges with an information content of less than 0.4 to improve matching accuracy. Since motif-finding procedures from our pipeline or others using importance scores are different from those generated by frequency-based methods, PWM representations varied and were required to be manually compared to known motifs in the final annotation step.

5.4 Results

5.4.1 Explain-seq predicts enhancer activity from DNA sequence

We designed a computational end-to-end pipeline, Explain-seq, to discover known and potential *de novo* motifs given genomic sequences (Fig 5.1A and Method). To learn the cis-regulatory code and grammar embedded in enhancer sequences, we have developed a CNN-based deep learning model to predict enhancer activity given DNA sequences (Fig 5.1B). We have downloaded and pre-processed 6 public STARR-seq datasets in ENCODE Phase 4 (Method). The public genome-wide enhancer activity maps provide high-quality datasets to build predictive models of enhancer activity in a cell-line-specific manner. To demonstrate Explain-seq usage, we only focused on the

A549 cell line to develop our model and pipeline. In total, we had 1,027,953 regions of size 499 bp on the hg38 genome reference for the A549 cell line. Classification-based and regression-based two CNN models have been deployed to map 499 bp DNA sequences to their enhancer activities (Fig 5.1B and Method). The two models' architectures are similar to that of the Beluga model except for the output layer. The regression model has only one output node in the final layer, while the multi-label classification model has 10 output nodes representing different activity ranges which are then followed by a SoftMax layer to determine the final prediction (Method). We transferred all the weights in convolutional layers from the well-trained Beluga model and fine-tuned all parameters when re-trained with STARR-seq datasets (Fig 5.1A). Transfer learning can speed up training when applied to related task model training, and improve accuracy. Our models' training and validation loss function converged within 20 epochs given a genome-wide training task (Fig 5.2A-D). After training, we accurately predicted the enhancer activity signals for both models (Fig 5.1C).

Our regression model outperformed the multi-label classification model and the gkm-SVM model when evaluated on test hold-out chromosomes, chr8 and chr9 (Fig 5.1D). The predicted and observed enhancer-activity profiles are similar with Pearson correlation coefficient (PCC) 0.31, 0.21, 0.25, for a regression model, multi-label classification model, and gkm-SVM model, respectively. Our regression model outperformed the multi-label classification model which indicates that ENCODE STARR-seq datasets' signal-to-noise ratios are high enough to be predicted in high resolution with respect to continuous values instead of categorical labels. Even the best-performance regression model is not close to the concordance between experimental replicates (PCC=0.96). It indicates the need for additional information, such as open chromatin and TF binding in the A549 cell line, rather than DNA sequences alone to further boost prediction accuracy. Chen et al. trained a model on the same A549 STARR-seq dataset with input as signal shape to predict binary binding [135]. Although they achieved validation AUROC 0.9984 and AUPR 0.9978 using a binary classification model with different preprocessing steps and input, their model was not cell-type specific since the signal shape was transferable between cell lines and was not able to be leveraged for predictive cell-type specific enhancer motifs. We trained a final regression model with both replicates to proceed with interpretation in our pipeline (Fig 5.2B & D).

5.4.2 Explain-seq reveals important TF motifs

Interpreting our regression model with Explain-seq reveals known TF motifs in the A549 cell line, and potential *de novo* motifs. Given the regression model, DeepLIFT calculated the importance scores of each nucleotide in the selected input sequence whose mean signal is larger than 3 (Fig 5.3A). To summarize all seqlets into readable motifs, TF-Modisco clustered 20,000 seqlets into consolidated potential motifs. Those motifs were then compared to motifs in the JASPAR database using STAMP to verify with known motifs. Briefly, A549 is the most used human non-small cell lung cancer cell line, consisting of hypotriploid alveolar basal epithelial cells. By applying STAMP, we searched for *de novo* motifs and compared them to known PWMs. Overall, we identified 47 known motifs including ZEB1, SP1, YY1, and INSM1 (Fig 5.3B). ZEB1, zinc-finger e-box binding 1, is part of the ZEB family in humans. ZEB1 is involved in the generation and maintenance of epithelial cell polarity and its expression in epithelial cells results in epithelial-mesenchymal transition (EMT) [146]. SP1, specificity protein 1, is important for lung cancer cell proliferation and metastasis during tumorigenesis [147]. Transcription factor Yin Yang 1 (YY1) is associated with the EMT process in the A549 cell line and regulates pulmonary fibrotic progression in lung epithelial cells [148]. INSM1, identified from STAMP in the A549 cell line, is a sensitive marker for the neuroendocrine differentiation of human lung tumors [149]. In summary, the determined known motifs are A549 cell line specific, indicating corresponding biological activities of the hypotriploid alveolar basal epithelial cells from the A549 cell line. Additionally, STAMP enables us to identify *de novo* motifs (Fig 5.3C). For example, some *de novo* motifs enrich for CpG sites. CpG sites are regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the sequence. CpG sites are highly related to DNA methylation that occurs more frequently by hypermethylation in cancers. Given the novel TF motifs, we are able to explore more biological insights within the A549 cell line.

5.5 Discussion

Deep learning algorithms like CNN have been widely used in DNA sequence analysis within the whole genome. However, deep learning model interpretation and biological explanation remain changeling. Here, we introduced a novel end-to-end computational pipeline, Explain-seq, to automate the process of developing and interpreting deep learning models in the context of genomics. We demonstrated the usage of Explain-seq with new STARR-seq datasets to predict enhancer activities, characterize cis-regulatory code, and identify known and *de novo* motifs using deep learning algorithms. Explain-seq quantitatively predicts enhancer signals in continuous values. Transfer learning is applied to improve the model accuracy and increase computational efficiency. Also, Explain-seq provided insights into the interpretation of the deep learning model and identify biological relevance. Specifically, Explain-seq reveals the sequence-to-function relationship by calculating nucleotide importance scores. Furthermore, by comparing with the existing motif database, Explain-seq successfully determines cell type-specific known and *de novo* motifs, which may contribute to the functionality.

Future work will aim to address the challenge of generalizing predictive models from a single cell line to multiple cell lines and different cell types across organisms. Almeida *et al.* [150] demonstrated the potential for generalization by developing DeepSTARR models trained on Drosophila enhancer data, which were then successfully transferred to humans. It is worth noting that both our models (the DeepSTARR and the CNN models trained with Explain-seq) share similar CNN architectures and interpretation suites, despite being developed independently and around the same time. We anticipate that by integrating various STARR-seq datasets with detailed biological interpretations, we will be able to decode gene regulatory information across entire genomes.

5.6 Availability

Explain-seq pipeline: `https://github.com/nsamzhao/Explain-seq`

Data at Zenodo: `10.5281/zenodo.7526380`

5.7 Publication

The study in this chapter is on *bioRxiv* [151]: Zhao, N., Wang, S., Huang, Q., Dong, S., & Boyle, A. P. (2023). Explain-seq: an end-to-end pipeline from training to interpretation of sequence-based deep learning models. *bioRxiv*, 2023-01.

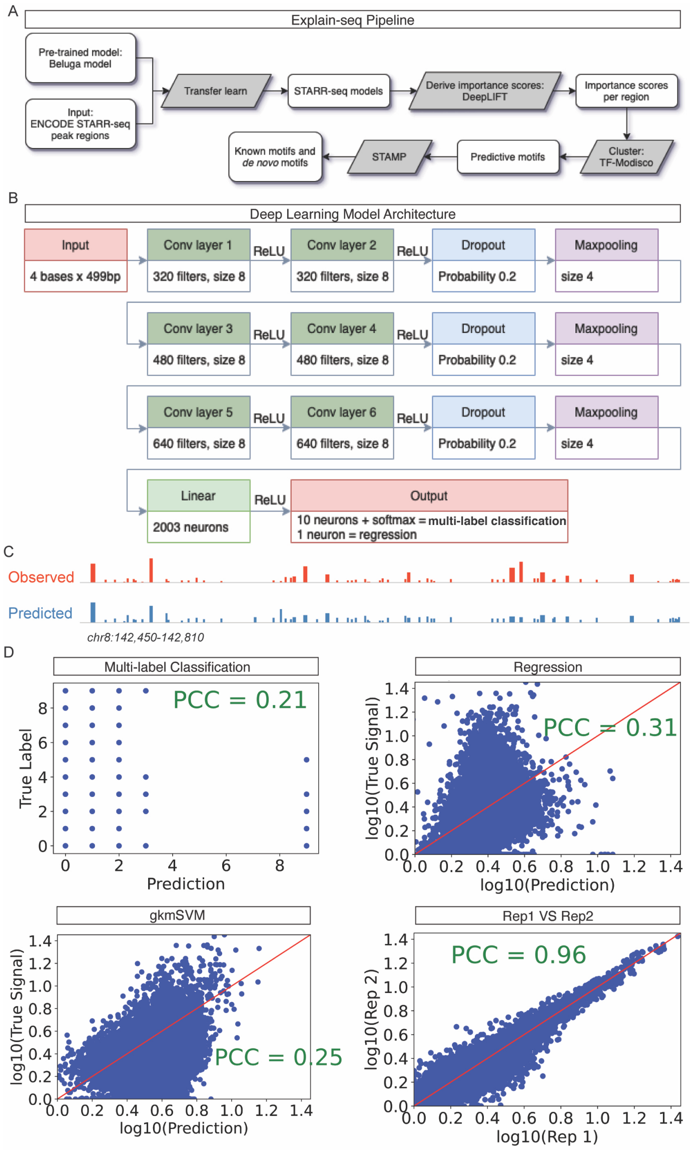


Figure 5.1: Explain-seq Predicts Enhancer Activity from Genome-wide DNA sequences. (A) Overview of Explain-seq pipeline to infer enhancer activities from the A549 cell line and to identify known and de novo motifs. (B) Architecture of the convolutional neural network (CNN) that was trained to predict A549 enhancer activities from 499bp sequences. Both the regression model and the multi-label classification model employ the same architecture. (C) Explain-seq predicts enhancer activity genome-wide. The IGV genome tracks screenshot depicts observed and predicted signals using the regression model for a locus on the held-out test chromosome 8. (D) Explain-seq with CNN regression model predicts enhancer activity better than gkm-SVM and the CNN multi-label classification model. Scatter plots of predicted vs. observed enhancer activity signals across all DNA sequences in the test set chromosomes are shown for CNN multi-label classification model, CNN regression model, and gkm-SVM. We also calculated the correspondence between the actual signals in one biological replicate of A549 with the actual signals in another biological replicate to serve as the maximum prediction accuracy threshold for this task.

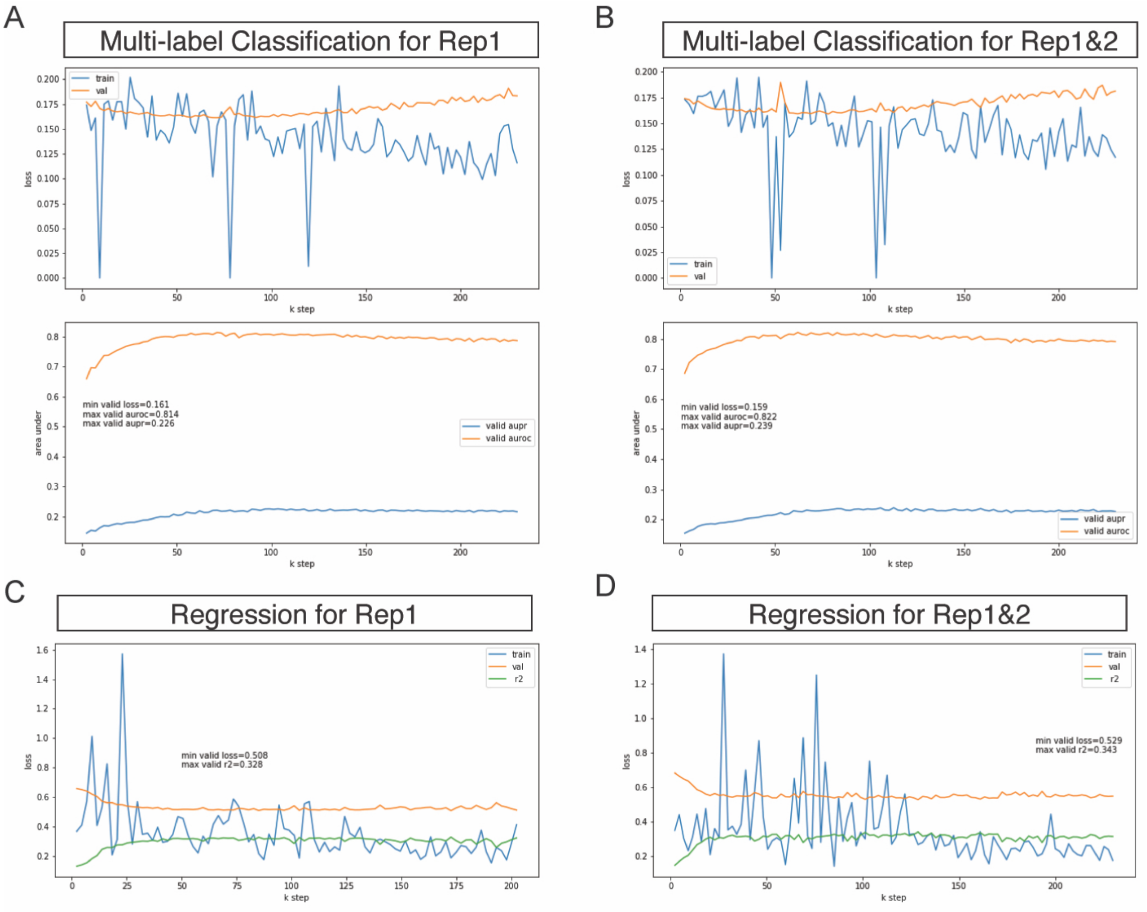


Figure 5.2: Multi-label Classification and Regression Model Performances. (A) Loss function along the iterations of multi-label classification model for replicate 1 (top row). Model performance on the test data using AUPR and AUROC metrics to measure (second row). (B) Loss function along the iterations of the multi-label classification model for replicate 1 and replicate 2 (top row). Model performance on the test data using AUPR and AUROC metrics (second metrics). (C) Loss function along the iterations of the regression model replicate 1. (D) Loss function along the iterations of the regression model replicate 1 and replicate 2.

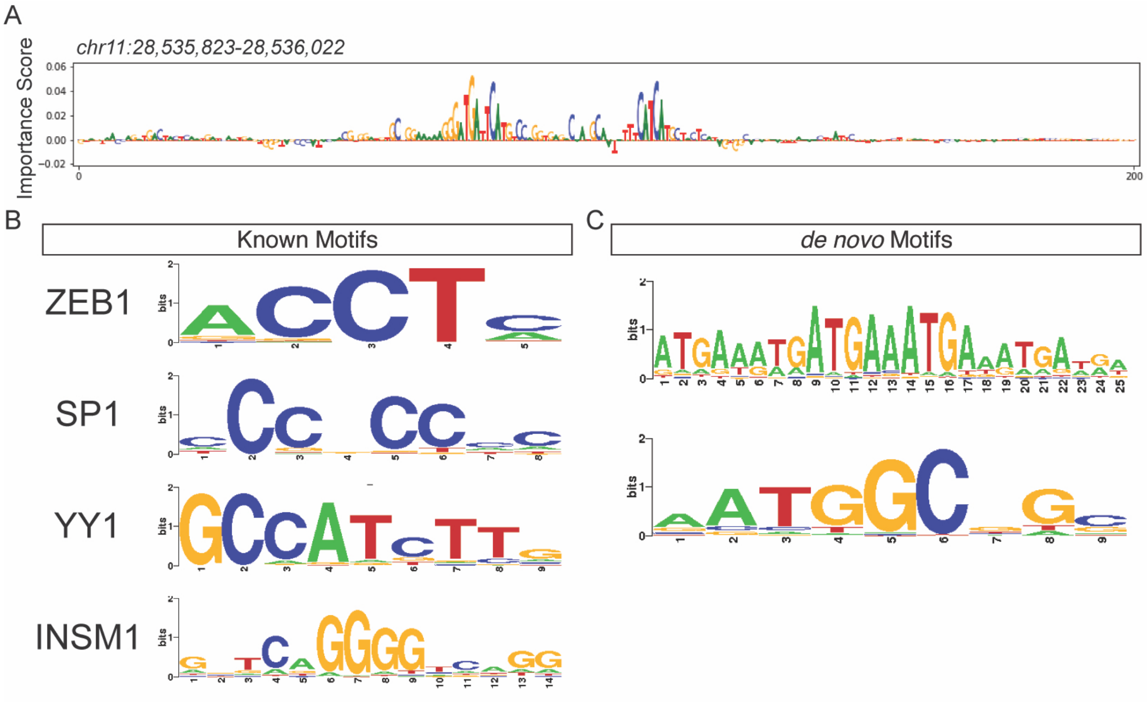


Figure 5.3: Explain-seq reveals known motifs and de novo motifs for enhancer activity. (A) Within Explain-seq, DeepLIFT calculated the importance scores of each base pair in selected input sequences whose mean signal is larger than 3. Important sub-sequences are highlighted. (B) Explain-seq identified known motifs (C) and Explain-seq revealed some de novo motifs for the A549 cell line.

CHAPTER VI

Conclusions and Future Directions

6.1 Summary

The main objective of this dissertation is to uncover biological insights from large-scale high-throughput sequencing genomic datasets, with a specific focus on non-coding regulatory mechanisms. To achieve this goal, I addressed four key questions: (1) how to process the raw high-throughput sequencing data into human-readable and processable data for further downstream analysis, (2) how to store those results and make them easily accessible to study, such as regulatory elements and variants, (3) how to prioritize regulatory variants using all available data, and (4) how to summarize and represent findings, such as into motifs.

With the massive datasets generated in genomics, I developed a series of computational methods and tools following the map outlined in Fig 6.1 to study non-coding regulatory elements and variants.

In Chapter II, I first developed a peak calling software, F-Seq2, to accurately identify the regulatory regions for common genomics assays. F-Seq2 outperformed the state-of-the-art models including MACS2 when evaluating on DNase-seq, ATAC-seq, and TF and histone ChIP-seq datasets in terms of precision and recall. It emphasized the importance of developing such low-level software, which determines

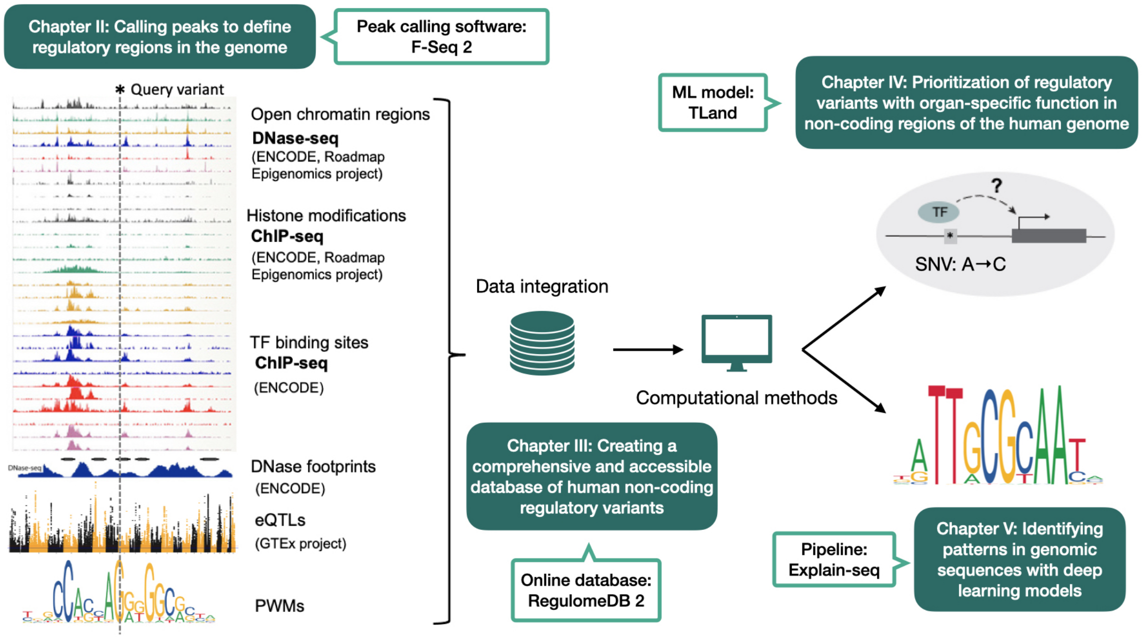

Chapter II: Calling peaks to define regulatory regions in the genome
Peak calling software: F-Seq 2
✱ Query variant
Open chromatin regions
DNase-seq
(ENCODE, Roadmap Epigenomics project)
Histone modifications
ChIP-seq
(ENCODE, Roadmap Epigenomics project)
TF binding sites
ChIP-seq
(ENCODE)
DNase-seq
DNase footprints
(ENCODE)
eQTLs
(GTEx project)
PWMs
Data integration
Computational methods
Chapter III: Creating a comprehensive and accessible database of human non-coding regulatory variants
Online database: RegulomeDB 2
ML model: TLand
Chapter IV: Prioritization of regulatory variants with organ-specific function in non-coding regions of the human genome
?
TF
SNV: A→C
Pipeline: Explain-seq
Chapter V: Identifying patterns in genomic sequences with deep learning models

Figure 6.1: Roadmap of developed computational methods and tools.

the accuracy of any downstream analysis to draw any biological conclusions.

In Chapter III, I advanced the leading non-coding regulatory variants database RegulomeDB to its second version. The new interface design enables users to easily access comprehensive information about regulatory variants from large consortia, such as ENCODE and RoadMap. RegulomeDB2 provides a summarization score that incorporates all the evidence it found during a query, which represents how likely a query variant is to be a regulatory variant given all the hits of functional genomics data. The database now includes five times more data than the previous version and includes the new GRCh38 assembly. In this chapter, I also demonstrated how to effectively use ReguloemDB2 to form testable hypotheses given the variants of interest.

In Chapter IV, I developed a machine learning model, TLand, to prioritize regulatory variants in an organ-specific manner. Generalizing models to new cell lines or organs is still a challenge in genomics. TLand can accurately generalize to unseen organs by taking advantage of RegulomeDB-derived features, and larger-scale of modeling (i.e. at the organ level than at the cell level). TLand prioritized the correct organs for around 2 million GWAS SNPs by taking into account data availability issues commonly existing in many sequence-based deep learning models.

In Chapter V, I created a pipeline, Explain-seq, to automate the process from the training of sequence-based deep learning models to interpreting predictive subpatterns into motifs. I demonstrated the usage by training on recent STARR-seq assays for enhancers. Known motifs compared to the JASPAR database, and *de novo* motifs were identified in a cell-specific manner.

6.2 Future directions

6.2.1 Integrating multiple data modalities and combining models to improve model performance

Sequence-based learning models have a potential upper ceiling of performance. As mentioned earlier in the introduction chapter, regulations such as cross-chromosome regulations cannot be modeled by only expanding the range of those sequence models. More types of data are needed to integrate into the final model such as higher-order structure information through 3D chromatin data Hi-C [112]. Due to the different data modalities (DNA sequences and Hi-C maps), integrating such data often involves late integration (i.e. model ensembling) or intermediate integrating (i.e. multi-modal learning, see both definitions in the introduction). However, the key principles of the optimal design to integrate higher-order data and other molecular interaction datasets are still not fully understood. More research is needed to be conducted in the area.

Combining a comprehensive set of models (e.g. models from the zoo of genomics, kipoi [152]) could potentially allow us to model higher-scale features, including an organism's phenotypes. However, although the cost and speed of high-throughput sequencing data have decreased dramatically, the speed of collecting phenotypic data has not progressed at the same pace. Previously, medical records were the preferred source of information for medical conditions, but emerging research is exploring internet and mobile technologies as viable approaches for large population phenotyping [54]. Compared to medical record reviews, internet-based phenotyping, such as self-reporting, can be performed quickly. Tung et al. [153] evaluated ¿ 20,000 individuals for 50 phenotypes, including Crohn's disease, inflammatory bowel disease, and diabetes in 1 year using only a small team. However, the quality of such phenotypic data should be evaluated carefully before training models.

6.2.2 Interpreting complex models

As machine learning models become more complex to capture complex dependencies in data, the interpretability of such models tends to decrease. One example of this is linear regression models, which have easily interpretable coefficients that can be used to understand the impact of individual variables on the outcome. However, deep learning models, such as neural networks, are more difficult to interpret due to their complex structure and a high number of parameters.

To address this issue, researchers have developed a variety of methods for interpreting deep learning models. One popular approach is to use visualization techniques to gain insight into the internal workings of the model. For example, activation maps can be used to visualize which parts of an image are important for a neural network's classification decision. Another approach is to use algorithms such as LIME (Local Interpretable Model-Agnostic Explanations) [154] or SHAP (SHapley Additive exPlanations) [155] to dissect what the model has learned. LIME generates "local" explanations for individual predictions by fitting a simpler, interpretable model to the local region around the prediction. SHAP values are another way of generating model-agnostic explanations that quantify the contribution of each feature to a particular prediction using game theory.

However, even with these methods, interpreting deep learning models can still be challenging. As a result, it is often recommended to use simpler models when possible, especially when interpretability is a priority. While complex models may offer improved performance, their lack of interpretability can make it difficult to understand how and why they are making decisions, which can be a major concern in fields such as healthcare where the stakes are high. It is a tradeoff between expressiveness and interpretability, and one needs to choose carefully according to the task.

6.2.3 Extending gold standards of ASB SNVs

Defining the gold standards of regulatory variants is challenging. We defined regulatory variants as non-coding variants having any regulatory functions. Since the delicate regulation of gene expression is achieved by the interplay between regulatory elements and TFs, any changes in TF binding to specific allele indicates the regulatory activities. Thus, we defined ASBs as our gold standard to train our model for regulatory variants. However, such ASB data is not comprehensive enough due to the limited TF ChIP-seq data for each cell line to call those ASBs. Moreover, the existing ASB data did not agree well with each other.

During the development of our TLand model, we attempted to improve its performance by ensembling more machine learning models, such as gkm-SVM but found that this did not yield better results. However, when we supplemented our ASB data with data from other sources, we saw improved performance. We hypothesized that our ASB datasets, despite being called using 600 TF ChIP-seq datasets, were not comprehensive enough. Furthermore, we only called ASBs in six cell lines, which may not be representative of all possible cell types. Recently, a new database called Adastra has been developed that contains 266,940 ASBs passing a 5% false discovery rate across 647 cell types and 1,043 TFs. We believe that training on such a comprehensive dataset could further improve our model's performance and enable it to become a transferable model for prioritizing regulatory variants across different organs.

6.2.4 Assigning prioritized variants to genes

The next goal of RegulomeDB is to assign the prioritized variants to genes. This is still an open question in genomics. To address this issue, researchers recently developed the activity-by-contact (ABC) model [156], which utilizes read counts of DHS and H3K27ac, as well as Hi-C maps as input. The ABC model is based on the biochemical concept that an enhancer's quantitative effect on a gene depends on its strength as an enhancer (Activity) weighed by how often it comes into 3D contact with the gene's promoter (Contact). The ABC model assumes that the relative contribution of an element on a gene's expression should depend on that element's effect divided by the total effect of all elements. This model provides a straightforward and precise way of linking enhancers to genes. To improve the performance of the ABC model, we hypothesize that substituting the enhancer activity scores with the RegulomeDB score, which considers all ENCODE assays including H3K27ac, could be beneficial. By doing so, we may integrate the ABC model into our database as a linking method, which could help better prioritize regulatory variants and provide a comprehensive map of affected genes and gene networks.

6.3 Concluding remarks

In this dissertation, I developed a series of computational methods and tools to study regulatory variants. I developed peak-calling software to accurately define regulatory regions. Then I advanced a leading regulatory variant database into its second version by integrating a new interface and the latest functional genomics data. I developed a machine learning model to annotate and prioritize regulatory variants in an organ-specific manner. Lastly, I created an automation pipeline from training sequence-based deep learning models to interpreting predictive patterns of the models to motifs. We hope these methods can have broad applications to help researchers to characterize both known and *de novo* regulatory variants, and ultimately have an impact on precision medicine and developing clinical therapies.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

[2] R. V. Broekema, O. B. Bakker, and I. H. Jonkers, "A practical view of fine-mapping and gene prioritization in the post-genome-wide association era," *Open Biol.*, vol. 10, p. 190221, Jan. 2020.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] N. Ouyang and A. P. Boyle, "TRACE: transcription factor footprinting using chromatin accessibility data and DNA sequence."

[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] 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.

[17] N. Zhao and A. P. Boyle, "F-Seq2: improving the feature density based peak caller with dynamic statistics," *NAR Genom Bioinform*, vol. 3, p. lqab012, Mar. 2021.

[18] Y. Guo, S. Mahony, and D. K. Gifford, "High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints," *PLoS Comput. Biol.*, vol. 8, p. e1002638, Aug. 2012.

[19] H. Xing, Y. Mo, W. Liao, and M. Q. Zhang, "Genome-wide localization of protein-DNA binding and histone modification by a bayesian change-point method with ChIP-seq data," *PLoS Comput. Biol.*, vol. 8, p. e1002613, July 2012.

[20] A. Harmanci, J. Rozowsky, and M. Gerstein, "MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework," *Genome Biol.*, vol. 15, no. 10, p. 474, 2014.

[21] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol.*, vol. 9, p. R137, Sept. 2008.

[22] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data," *Bioinformatics*, vol. 25, pp. 1952–1958, Aug. 2009.

[23] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao, "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data," *Nucleic Acids Res.*, vol. 36, pp. 5221–5231, Sept. 2008.

[24] A. P. Boyle, J. Guinney, G. E. Crawford, and T. S. Furey, "F-Seq: a feature density estimator for high-throughput sequence tags," *Bioinformatics*, vol. 24, pp. 2537–2538, Nov. 2008.

[25] S. John, P. J. Sabo, R. E. Thurman, M.-H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos, "Chromatin accessibility pre-determines glucocorticoid receptor binding patterns," *Nat. Genet.*, vol. 43, pp. 264–268, Mar. 2011.

[26] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park, “Design and analysis of ChIP-seq experiments for DNA-binding proteins,” *Nat. Biotechnol.*, vol. 26, pp. 1351–1359, Dec. 2008.

[27] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong, “An integrated software system for analyzing ChIP-chip and ChIP-seq data,” *Nat. Biotechnol.*, vol. 26, pp. 1293–1300, Nov. 2008.

[28] X.-Y. Li, S. Thomas, P. J. Sabo, M. B. Eisen, J. A. Stamatoyannopoulos, and M. D. Biggin, “The role of chromatin accessibility in directing the widespread, overlapping patterns of drosophila transcription factor binding,” *Genome Biol.*, vol. 12, p. R34, Apr. 2011.

[29] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow, “Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data,” *Nat. Methods*, vol. 5, pp. 829–834, Sept. 2008.

[30] H. Wu and H. Ji, “PolyaPeak: detecting transcription factor binding sites from ChIP-seq using peak shape information,” *PLoS One*, vol. 9, p. e89694, Mar. 2014.

[31] M. Micsinai, F. Parisi, F. Strino, P. Asp, B. D. Dynlacht, and Y. Kluger, “Picking ChIP-seq peak detectors for analyzing chromatin modification experiments,” *Nucleic Acids Res.*, vol. 40, p. e70, May 2012.

[32] R. Thomas, S. Thomas, A. K. Holloway, and K. S. Pollard, “Features that define the best ChIP-seq peak calling algorithms,” *Brief. Bioinform.*, vol. 18, pp. 441–450, May 2017.

[33] 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.

[34] IGVF Consortium, “Impact of genomic variation on function (IGVF) consortium.” `https://igvf.org/`. Accessed: 2023-3-27.

[35] H. Zhou, T. Arapoglou, X. Li, Z. Li, X. Zheng, J. Moore, A. Asok, S. Kumar, E. E. Blue, S. Buyske, N. Cox, A. Felsenfeld, M. Gerstein, E. Kenny, B. Li, T. Matise, A. Philippakis, H. L. Rehm, H. J. Sofia, G. Snyder, NHGRI Genome Sequencing Program Variant Functional Annotation Working Group, Z. Weng, B. Neale, S. R. Sunyaev, and X. Lin, “FAVOR: functional annotation of variants online resource and annotator for variation across the human genome,” *Nucleic Acids Res.*, vol. 51, pp. D1300–D1311, Jan. 2023.

[36] 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.

[37] X. Li, Z. Li, H. Zhou, S. M. Gaynor, Y. Liu, H. Chen, R. Sun, R. Dey, D. K. Arnett, S. Aslibekyan, and Others, “Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale,” *Nat. Genet.*, vol. 52, no. 9, pp. 969–983, 2020.

[38] 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.

[39] S. Dong and A. P. Boyle, “Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome,” *Nucleic Acids Res.*, vol. 50, p. e6, Jan. 2022.

[40] M. Pazin, “Genomics of gene regulation.” `https://www.genome.gov/Funded-Programs-Projects/Genomics-of-Gene-Regulation`. Accessed: 2023-3-27.

[41] D. Lee, D. U. Gorkin, M. Baker, B. J. Strober, A. L. Asoni, A. S. McCallion, and M. A. Beer, “A method to predict the impact of regulatory variants from DNA sequence,” *Nat. Genet.*, vol. 47, pp. 955–961, Aug. 2015.

[42] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, “Deep learning: new computational modelling techniques for genomics,” *Nat. Rev. Genet.*, vol. 20, pp. 389–403, July 2019.

[43] P. D’haeseleer, “What are DNA sequence motifs?,” *Nat. Biotechnol.*, vol. 24, pp. 423–425, Apr. 2006.

[44] J. Zhou and O. G. Troyanskaya, “Predicting effects of noncoding variants with deep learning–based sequence model,” *Nat. Methods*, vol. 12, pp. 931–934, Aug. 2015.

[45] D. R. Kelley, J. Snoek, and J. L. Rinn, “Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks,” *Genome Res.*, vol. 26, pp. 990–999, July 2016.

[46] 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.

[47] K. M. Chen, A. K. Wong, O. G. Troyanskaya, and J. Zhou, “A sequence-based global map of regulatory activity for deciphering human genetics,” *Nat. Genet.*, vol. 54, pp. 940–949, July 2022.

[48] 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.

[49] D. Quang and X. Xie, “FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data,” *Methods*, vol. 166, pp. 40–47, Aug. 2019.

[50] S. Bai, J. Zico Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” Mar. 2018.

[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” June 2017.

[52] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, “Effective gene expression prediction from sequence by integrating long-range interactions,” *Nat. Methods*, vol. 18, pp. 1196–1203, Oct. 2021.

[53] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, “DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome,” *Bioinformatics*, Feb. 2021.

[54] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities,” *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019.

[55] D. H. Wolpert, “Stacked generalization,” *Neural Netw.*, vol. 5, pp. 241–259, Jan. 1992.

[56] 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.

[57] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb, “FAIRE (Formaldehyde-Assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin,” *Genome Res.*, vol. 17, pp. 877–885, June 2007.

[58] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing,” *Nat. Methods*, vol. 4, pp. 651–657, Aug. 2007.

[59] M. Rosenblatt, “Remarks on some nonparametric estimates of a density function,” *Ann. Math. Stat.*, vol. 27, no. 3, pp. 832–837, 1956.

[60] E. Parzen, “On estimation of a probability density function and mode,” *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962.

[61] P. Ramachandran and T. J. Perkins, “Adaptive bandwidth kernel density estimation for next-generation sequencing data,” *BMC Proc.*, vol. 7, p. S7, Dec. 2013.

[62] H. Koohy, T. A. Down, M. Spivakov, and T. Hubbard, “A comparison of peak callers used for DNase-Seq data,” *PLoS One*, vol. 9, p. e96303, May 2014.

[63] N. Hiranuma, S. M. Lundberg, and S.-I. Lee, “AIControl: replacing matched control experiments with machine learning improves ChIP-seq peak identification,” *Nucleic Acids Res.*, vol. 47, p. e58, June 2019.

[64] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel, “Measuring reproducibility of high-throughput experiments,” *aoas*, vol. 5, pp. 1752–1779, Sept. 2011.

[65] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *J. R. Stat. Soc.*, vol. 57, pp. 289–300, Jan. 1995.

[66] 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.

[67] T. Hastie and R. Tibshirani, “Generalized additive models: Some applications,” *J. Am. Stat. Assoc.*, vol. 82, pp. 371–386, June 1987.

[68] H. Touzet and J.-S. Varré, “Efficient and accurate p-value computation for position weight matrices,” *Algorithms Mol. Biol.*, vol. 2, p. 15, Dec. 2007.

[69] 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.

[70] E. D. Tarbell and T. Liu, “HMMRATAC: a hidden markov ModeleR for ATAC-seq,” *Nucleic Acids Res.*, vol. 47, pp. e91–e91, June 2019.

[71] M. Karimzadeh and M. M. Hoffman, “Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome.” Mar. 2019.

[72] A. Awdeh, M. Turcotte, and T. J. Perkins, “WACS: Improving ChIP-seq peak calling by optimally weighting controls.” Nov. 2019.

[73] N. Zhao and A. P. Boyle, “F-seq2: improving the feature density based peak caller with dynamic statistics,” *NAR Genomics and Bioinformatics*, vol. 3, no. 1, p. lqab012, 2021.

[74] 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.

[75] 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.

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

[77] 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.

[78] D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S.-B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y.-D. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Köttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zöllner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, and G. R. Abecasis, "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program," *Nature*, vol. 590, pp. 290–299, Feb. 2021.

[79] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, and

D. G. MacArthur, "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature*, vol. 581, pp. 434–443, May 2020.

[80] 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.

[81] 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.

[82] H. Touzet and J.-S. Varré, "Efficient and accurate p-value computation for position weight matrices," *Algorithms Mol. Biol.*, vol. 2, p. 15, Dec. 2007.

[83] 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.

[84] 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.

[85] C. A. Boix, B. T. James, Y. P. Park, W. Meuleman, and M. Kellis, "Regulatory genomic circuitry of human disease loci by integrative epigenomics," *Nature*, vol. 590, pp. 300–307, Feb. 2021.

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

[87] 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.

[88] J. Schwartzentruber, S. Foskolou, H. Kilpinen, J. Rodrigues, K. Alasoo, A. J. Knights, M. Patel, A. Goncalves, R. Ferreira, C. L. Benn, A. Wilbrey, M. Bictash, E. Impey, L. Cao, S. Lainez, A. J. Loucif, P. J. Whiting, HIPSCI Consortium, A. Gutteridge, and D. J. Gaffney, "Molecular and functional variation in iPSC-derived sensory neurons," *Nat. Genet.*, vol. 50, pp. 54–61, Jan. 2018.

[89] S. Khetan, R. Kursawe, A. Youn, N. Lawlor, A. Jillette, E. J. Marquez, D. Ucar, and M. L. Stitzel, "Type 2 Diabetes–Associated genetic variants regulate chromatin accessibility in human islets," *Diabetes*, vol. 67, pp. 2466–2477, Sept. 2018.

[90] R. E. Gate, C. S. Cheng, A. P. Aiden, A. Siba, M. Tabaka, D. Lituiev, I. Machol, M. G. Gordon, M. Subramaniam, M. Shamim, K. L. Hougen, I. Wortman, S.-C. Huang, N. C. Durand, T. Feng, P. L. De Jager, H. Y. Chang, E. L. Aiden, C. Benoist, M. A. Beer, C. J. Ye, and A. Regev, "Genetic determinants of co-accessible chromatin regions in activated T cells across humans," *Nat. Genet.*, vol. 50, pp. 1140–1150, Aug. 2018.

[91] 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.

[92] 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.

[93] Q. Zhao, M. Dacre, T. Nguyen, M. Pjanic, B. Liu, D. Iyer, P. Cheng, R. Wirka, J. B. Kim, H. B. Fraser, and T. Quertermous, "Molecular mechanisms of coronary disease revealed using quantitative trait loci for TCF21 binding, chromatin accessibility, and chromosomal looping," *Genome Biol.*, vol. 21, p. 135, June 2020.

[94] D. Liang, A. L. Elwell, N. Aygün, O. Krupa, J. M. Wolter, F. A. Kyere, M. J. Lafferty, K. E. Cheek, K. P. Courtney, M. Yusupova, M. E. Garrett, A. Ashley-Koch, G. E. Crawford, M. I. Love, L. de la Torre-Ubieta, D. H. Geschwind, and J. L. Stein, "Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation," *Nat. Neurosci.*, vol. 24, pp. 941–953, July 2021.

[95] K. W. Currin, M. R. Erdos, N. Narisu, V. Rai, S. Vadlamudi, H. J. Perrin, J. R. Idol, T. Yan, R. D. Albanus, K. A. Broadaway, A. S. Etheridge, L. L. Bonnycastle, P. Orchard, J. P. Didion, A. S. Chaudhry, NISC Comparative Sequencing Program, F. Innocenti, E. G. Schuetz, L. J. Scott, S. C. J. Parker, F. S. Collins, and K. L. Mohlke, "Genetic effects on liver chromatin accessibility identify disease regulatory variants," *Am. J. Hum. Genet.*, vol. 108, pp. 1169–1189, July 2021.

[96] R. M. Kuhn, D. Haussler, and W. J. Kent, "The UCSC genome browser and associated tools," *Brief. Bioinform.*, vol. 14, pp. 144–161, Mar. 2013.

[97] S. Dong and A. P. Boyle, "Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome," *Nucleic Acids Res.*, Oct. 2021.

[98] B. Brocke, K.-P. Lesch, D. Armbruster, D. A. Moser, A. Müller, A. Strobel, and C. Kirschbaum, "Stathmin, a gene regulating neural plasticity, affects fear and anxiety processing in humans," *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, vol. 153B, pp. 243–251, Jan. 2010.

[99] J. D. Hocker, O. B. Poirion, F. Zhu, J. Buchanan, K. Zhang, J. Chiou, T.-M. Wang, Q. Zhang, X. Hou, Y. E. Li, Y. Zhang, E. N. Farah, A. Wang, A. D. McCulloch, K. J. Gaulton, B. Ren, N. C. Chi, and S. Preissl, "Cardiac cell type–specific gene regulatory programs and disease risk association," *Science Advances*, vol. 7, no. 20, p. eabf1444, 2021.

[100] J. E. Horowitz, J. A. Kosmicki, A. Damask, D. Sharma, G. H. L. Roberts, A. E. Justice, N. Banerjee, M. V. Coignet, A. Yadav, J. B. Leader, A. Marcketta, D. S. Park, R. Lanche, E. Maxwell, S. C. Knight, X. Bai, H. Guturu, D. Sun, A. Baltzell, F. S. P. Kury, J. D. Backman, A. R. Girshick, C. O'Dushlaine, S. R. McCurdy, R. Partha, A. J. Mansfield, D. A. Turissini, A. H. Li, M. Zhang, J. Mbatchou, K. Watanabe, L. Gurski, S. E. McCarthy, H. M. Kang, L. Dobbyn, E. Stahl, A. Verma, G. Sirugo, Regeneron Genetics Center, M. D. Ritchie, M. Jones, S. Balasubramanian, K. Siminovitch, W. J. Salerno, A. R. Shuldiner, D. J. Rader, T. Mirshahi, A. E. Locke, J. Marchini, J. D. Overton, D. J. Carey, L. Habegger, M. N. Cantor, K. A. Rand, E. L. Hong, J. G. Reid, C. A. Ball, A. Baras, G. R. Abecasis, and M. A. R. Ferreira, "Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19

risk and yields risk scores associated with severe disease," *Nat. Genet.*, vol. 54, pp. 382–392, Apr. 2022.

[101] T. H. Beacon, G. P. Delcuve, and J. R. Davie, "Epigenetic regulation of ACE2, the receptor of the SARS-CoV-2 virus1," *Genome*, vol. 64, pp. 386–399, Apr. 2021.

[102] N. Kubota and M. Suyama, "An integrated analysis of public genomic data unveils a possible functional mechanism of psoriasis risk via a long-range ERRFI1 enhancer," *BMC Med. Genomics*, vol. 13, p. 8, Jan. 2020.

[103] S. Dong, N. Zhao, E. Spragins, M. S. Kagda, M. Li, P. Assis, O. Jolanki, Y. Luo, J. M. Cherry, A. P. Boyle, and B. C. Hitz, "Annotating and prioritizing human non-coding variants with RegulomeDB v.2," *Nat. Genet.*, pp. 1–3, Apr. 2023.

[104] 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.

[105] 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.

[106] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nat. Methods*, vol. 4, pp. 651–657, Aug. 2007.

[107] 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.

[108] 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.

[109] International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, pp. 1299–1320, Oct. 2005.

[110] International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Q.

Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart, "A second generation human haplotype map of over 3.1 million SNPs," *Nature*, vol. 449, pp. 851–861, Oct. 2007.

[111] M. Pazin, D. A. Gilchrist, and S. A. Morris, "Impact of genomic variation on function (IGVF) consortium." `https://www.genome.gov/Funded-Programs-Projects/Impact-of-Genomic-Variation-on-Function-Consortium`. Accessed: 2023-4-2.

[112] 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.

[113] 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.

[114] S. Abramov, A. Boytsov, D. Bykova, D. D. Penzar, I. Yevshin, S. K. Kolmykov, M. V. Fridman, A. V. Favorov, I. E. Vorontsov, E. Baulin, F. Kolpakov, V. J. Makeev, and I. V. Kulakovskiy, "Landscape of allele-specific transcription factor binding in the human genome," *Nat. Commun.*, vol. 12, p. 2751, May 2021.

[115] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[116] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282 vol.1, Aug. 1995.

[117] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, Dec. 1943.

[118] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), pp. 2623–2631, Association for Computing Machinery, July 2019.

[119] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," pp. 2825–2830, Jan. 2012.

[120] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack," *J. Open Source Softw.*, vol. 3, p. 638, Apr. 2018.

[121] W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik, "BigWig and BigBed: enabling browsing of large distributed datasets," *Bioinformatics*, vol. 26, pp. 2204–2207, Sept. 2010.

[122] S. Neph, M. Scott Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E. Rynes, M. T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, and J. A. Stamatoyannopoulos, "BEDOPS: high-performance genomic feature operations," 2012.

[123] M. van der Sande and S. van Heeringen, "qnorm," June 2021.

[124] "ChatGPT." `https://chat.openai.com/`. Accessed: 2023-4-2.

[125] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.

[126] K. K. Dey, B. van de Geijn, S. S. Kim, F. Hormozdiari, D. R. Kelley, and A. L. Price, "Evaluating the informativeness of deep learning annotations for human complex diseases," *Nat. Commun.*, vol. 11, p. 4703, Sept. 2020.

[127] K. K. Dey, S. S. Kim, S. Gazal, J. Nasser, J. M. Engreitz, and A. L. Price, "Integrative approaches to improve the informativeness of deep learning models for human complex diseases." Aug. 2021.

[128] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *jair*, vol. 10, pp. 271–289, May 1999.

[129] W. Zhu, X. He, K. Cheng, L. Zhang, D. Chen, X. Wang, G. Qiu, X. Cao, and X. Weng, "Ankylosing spondylitis: etiology, pathogenesis, and treatments," *Bone research*, vol. 7, no. 1, p. 22, 2019.

[130] 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.

[131] C. D. Arnold, D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark, "Genome-wide quantitative enhancer activity maps identified by STARR-seq," *Science*, vol. 339, pp. 1074–1077, Mar. 2013.

[132] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, and J. Zeitlinger, "Base-resolution models of transcription-factor binding reveal soft motif syntax," *Nat. Genet.*, vol. 53, pp. 354–366, Mar. 2021.

[133] 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.

[134] M. Ghandi, M. Mohammad-Noori, N. Ghareghani, D. Lee, L. Garraway, and M. A. Beer, "gkmSVM: an R package for gapped-kmer SVM," *Bioinformatics*, vol. 32, pp. 2205–2207, July 2016.

[135] Z. Chen, J. Zhang, J. Liu, Y. Dai, D. Lee, M. R. Min, M. Xu, and M. Gerstein, "DECODE: a deep-learning framework for condensing enhancers and refining boundaries with large-scale functional assays," *Bioinformatics*, vol. 37, pp. i280–i288, July 2021.

[136] H. M. Amemiya, A. Kundaje, and A. P. Boyle, "The ENCODE blacklist: Identification of problematic regions of the genome," *Sci. Rep.*, vol. 9, p. 9354, June 2019.

[137] K. M. Chen, E. M. Cofer, J. Zhou, and O. G. Troyanskaya, "Selene: a PyTorch-based deep learning library for sequence data," *Nat. Methods*, vol. 16, pp. 315–318, Apr. 2019.

[138] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," Dec. 2019.

[139] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153, PMLR, 2017.

[140] A. Shrikumar, K. Tian, Ž. Avsec, A. Shcherbina, A. Banerjee, M. Sharmin, S. Nair, and A. Kundaje, "Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5," Oct. 2018.

[141] S. Mahony and P. V. Benos, "STAMP: a web tool for exploring DNA-binding motif similarities," *Nucleic Acids Res.*, vol. 35, pp. W253–8, July 2007.

[142] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk," *Nat. Genet.*, vol. 50, pp. 1171–1179, Aug. 2018.

[143] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014.

[144] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," Nov. 2017.

[145] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Res.*, vol. 37, pp. W202–8, July 2009.

[146] I. Georgakopoulos-Soares, D. V. Chartoumpekis, V. Kyriazopoulou, and A. Zaravinos, "EMT factors and metabolic pathways in cancer," *Front. Oncol.*, vol. 10, p. 499, Apr. 2020.

[147] T.-I. Hsu, M.-C. Wang, S.-Y. Chen, Y.-M. Yeh, W.-C. Su, W.-C. Chang, and J.-J. Hung, "Sp1 expression regulates lung tumor progression," *Oncogene*, vol. 31, pp. 3973–3988, Aug. 2012.

[148] C. Zhang, X. Zhu, Y. Hua, Q. Zhao, K. Wang, L. Zhen, G. Wang, J. Lü, A. Luo, W. C. Cho, X. Lin, and Z. Yu, "YY1 mediates TGF-β1-induced EMT and pro-fibrogenesis in alveolar epithelial cells," *Respir. Res.*, vol. 20, p. 249, Nov. 2019.

[149] J. N. Rosenbaum, Z. Guo, R. M. Baus, H. Werner, W. M. Rehrauer, and R. V. Lloyd, "INSM1: A novel immunohistochemical and molecular marker for neuroendocrine and neuroepithelial neoplasms," *Am. J. Clin. Pathol.*, vol. 144, pp. 579–591, Oct. 2015.

[150] B. P. de Almeida, F. Reiter, M. Pagani, and A. Stark, "DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers," *Nat. Genet.*, vol. 54, pp. 613–624, May 2022.

[151] N. Zhao, S. Wang, Q. Huang, S. Dong, and A. P. Boyle, “Explain-seq: an end-to-end pipeline from training to interpretation of sequence-based deep learning models.” Jan. 2023.

[152] Ž. Avsec, R. Kreuzhuber, J. Israeli, N. Xu, J. Cheng, A. Shrikumar, A. Banerjee, D. S. Kim, T. Beier, L. Urban, A. Kundaje, O. Stegle, and J. Gagneur, “The kipoi repository accelerates community exchange and reuse of predictive models for genomics,” *Nat. Biotechnol.*, vol. 37, pp. 592–600, June 2019.

[153] J. Y. Tung, C. B. Do, D. A. Hinds, A. K. Kiefer, J. M. Macpherson, A. B. Chowdry, U. Francke, B. T. Naughton, J. L. Mountain, A. Wojcicki, and N. Eriksson, “Efficient replication of over 180 genetic associations with self-reported medical data,” *PLoS One*, vol. 6, p. e23473, Aug. 2011.

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

[155] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” May 2017.

[156] C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, and J. M. Engreitz, “Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations,” *Nat. Genet.*, vol. 51, pp. 1664–1669, Nov. 2019.