Fine Mapping Functional Noncoding Genetic Elements Via Machine Learning

by

Pramod Bharadwaj Chandrashekar

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2020 by the
Graduate Supervisory Committee:

Li Liu, Chair
George Runger
Valentin Dinu

ARIZONA STATE UNIVERSITY

December 2020

ABSTRACT

All biological processes like cell growth, cell differentiation, development, and aging requires a series of steps which are characterized by gene regulation. Studies have shown that gene regulation is the key to various traits and diseases. Various factors affect the gene regulation which includes genetic signals, epigenetic tracks, genetic variants, etc. Deciphering and cataloging these functional genetic elements in the non-coding regions of the genome is one of the biggest challenges in precision medicine and genetic research. This thesis presents two different approaches to identifying these elements: TreeMap and DeepCORE. The first approach involves identifying putative causal genetic variants in cis-eQTL accounting for multisite effects and genetic linkage at a locus. TreeMap performs an organized search for individual and multiple causal variants using a tree guided nested machine learning method. DeepCORE on the other hand explores novel deep learning techniques that models the relationship between genetic, epigenetic and transcriptional patterns across tissues and cell lines and identifies co-operative regulatory elements that affect gene regulation. These two methods are believed to be the link for genotype-phenotype association and a necessary step to explaining various complex diseases and missing heritability.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to all members of my committee, Dr. Li Liu, Dr. Valentin Dinu, and Dr. Runger for their support and guidance throughout my PhD journey. I am deeply indebted to my advisor and mentor Dr. Li Liu for all her support throughout my PhD career. Dr. Liu has been a constant source of encouragement to me and has supported and believed in me throughout my journey. She has provided me extensive personal and professional guidance and has guided me to become a better researcher. I would also like to extend my deepest gratitude to Dr. Gerorge Runger who has provided continued support and advice on my PhD research. I am also extremely grateful to Dr. Valentin Dinu for providing me an opportunity to collaborate with him and for his invaluable guidance and insightful remarks on my dissertation.

I would like to thank the Biomedical Informatics program for introducing me to the world of informatics in healthcare. I would like to extend my sincere thanks to Dr. Graciela Gonzalez for her constant support and guidance since our class project in 2015. I would have not pursued my PhD journey if it weren't for her encouragement and guidance. I am also grateful to her for allowing me to work on several projects at UPenn with her. Many thanks to Lauren Madjidi and Maria Hanlin for their constant help and support since the first day.

I have been very fortunate to have started my PhD journey at BMI with my dearest friend, Arjun Magge. I can't thank him enough for all his help, brainstorming sessions, and friendly conversations. Special thanks to my labmate Navid Ahmadinejad for his support and stress-busting sessions. I am also thankful to Matthew Lee for helping me in formulating and running various research experiments. Many thanks to other members of my lab mates Abdulkadir Elmas, Brian Hanratty, and Anna Freydenzon. I

would also like to thank all my other PhD cohort Verah Nyarige, Akshay Vankipuram, Meredith Abrams, Lu Zheng, and Hiral Soni for being wonderful companions along this journey. I would also like to mention my thanks to my neighboring lab mates Ziqi, Juan, and Abigail for constantly checking in with each other, tea-time discussions, and numerous racquetball session.

My sincere thanks and deepest respect to my parents for their endless love, support, motivation, and sacrifices. They are my ultimate role models. I am also extremely thankful to my grandparents who are more worried about me than anyone else in this world. I would also thank all my family members who have supported me through the highs and lows of my life. Special thanks to my two uncles Kumar and Suresh for taking such great care of me, rooting for me all through, and constantly guiding through different phases of my life. I would also like to thank all my dear friends especially Abhishek, Gurudutta, Rohini, and Varsha for making this journey memorable.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1 INTRODUCTION 1

1.1 List of Publications 1

1.2 Overview of Gene Regulatory Mechanism 2

1.2.1 Cis-regulatory Elements 2

1.2.2 Trans-regulatory Elements 5

1.3 Genotype-Phenotype Association 8

1.3.1 Genetic Mutations 8

1.3.2 GWAS and eQTL Analysis 11

1.4 Computational Methods for Genotype-Phenotype Studies 12

1.4.1 Computational Methods to Discover DAVs and eQTL Variants 12

1.4.2 Computational Methods for Predicting the Functional Impact of Genetic Variants 13

1.4.3 Machine Learning and Deep Learning in Bioinformatics 19

1.5 Conclusion 22

2 TREEMAP: A STRUCTURED APPROACH TO FINE MAPPING OF EQTL VARIANTS 23

2.1 Background 23

2.2 Materials and Methods 25

2.2.1 Data Structure 25

2.2.2 TreeMap Framework 27

2.2.3 Simulation Data 30

CHAPTER Page

2.2.4 GTEx Datasets 31
2.3 Results 31
2.3.1 Performance on Simulation Data 32
2.3.2 Applications to GTEx data 36
2.4 Discussion 37
2.5 Limitations and Future Direction 39
3 DEEPCORE: INTERPRETABLE MULTI-VIEW DEEP LEARNING APPROACH FOR DETECTING CO-OPERATIVE REGULATORY ELEMENTS 41
3.1 Background 41
3.1.1 Gene Expression is a Molecular Phenotype 41
3.1.2 Extracting Useful Features from Deep Learning Models 46
3.2 The Novel DeepCORE Framework 48
3.2.1 Data Extraction and Preprocessing 48
3.2.2 DeepCORE Model for Tissue-specific Gene Expression Prediction 49
3.2.3 Detect Co-operative Regulatory Elements (CORES) 58
3.2.4 Training Procedure 59
3.3 Results and Discussion 60
3.3.1 Multi-view Architecture Improves the Accuracy of Gene Expression Prediction 60
3.4 Limitations and Future Direction 76
4 CONCLUSION AND FUTURE WORK 78
4.1 Integrating TreeMap and DeepCORE 79

CHAPTER Page

4.2 Extending DeepCORE for EWAS 79

REFERENCES 81

LIST OF TABLES

Table		Page
1.1	Different Classes of Histone Modifications (Kouzarides, 2007)	7
1.2	Types of Mutations Based on Functional Impact	9
1.3	Classes of DNA Mutations and Their Impact	10
1.4	Properties of Predictive Models for Six Tools	15
3.1	Different Hyperparameter Settings	61
3.2	Performance Comparison of Various Models on E061 and E071 Cell Lines.	63
3.3	Count of Histone Marks in the Attended Regions of High and Low Expression Genes.	68

LIST OF FIGURES

Figure Page

1.1 Cis-regulatory Elements (Modified from Maston *et al.* (2006) 3

1.2 Transcription Factors Binding to the Regulatory Regions Initiating Transcription (Blackwood and Kadonaga, 1998) 6

1.3 Histone Modifications Affecting Gene Regulation. (Top) Actively Transcribed Genes Are Marked by H3k27ac and H3k4me3. (Middle) This Represents a Poised/Bivalent State Where We Have Both Activators (H3k4me1) in the Enhancer Region and Repressors(H3k9me3, H3k7me3) in the Promoters Leading to Repression of Genes. (Bottom) Presence of H3k27me3 in Both Enhancers and Promoters Always Leads to Gene Repression. 8

1.4 Types of Genetic Mutations 9

1.5 Spectrum of Genetic Variants. 11

1.6 Performance of Tested Methods on Detecting Pathogenic Variants in Position-matched Ncsnvs. (a) ROC Curves with AUROC Values Displayed for Each Method. (b) Cumulative Distribution of CADD Scores. (c) Cumulative Distribution of DeepSEA Scores. 16

1.7 Performance of Tested Methods on Region-matched Pathogenic and Non-pathogenic ncSNVs. (a) Violin Plots Show Distributions of Impact Score Difference Between Nearby Pathogenic and Non-pathogenic ncSNVs. (b) Correlation of Impact Scores for ncSNVs Located Within given Genomic Distances. (c) Success Rates of Ranking a Pathogenic ncSNVs Higher than a Non-pathogenic ncSNVs Located Within 1000bps Vicinity. (d) Success Rates Stratified by the Genomic Context of Pathogenic ncSNVs. 18

Table Page

2.1 The TreeMap Method 26
2.2 Performance on Mapping Independent Causal Variants by TreeMap, DAP, Stepwise Conditional Analysis, CaVEMaN, and CAVIARBF. * Indicate Significant Differences of Performance Between TreeMap and Corresponding Methods (P< 0.05). 33
2.3 Performance on Mapping Independent Causal Variants by TreeMap, DAP, Stepwise Conditional Analysis, CaVEMaN, and CAVIARBF. 34
2.4 Precision-Recall Curves of TreeMap Using DHS-derived Functional Scores. 35
2.5 Average CPU Time (Seconds in Log Scale) to Analyze One Simulated Case. 36
2.6 Fractions of Genes with Single or Multiple eQTL Identified by TreeMap, TreeMap Restricted to 1Mb Flanking TSS, DAP, and CaVEMaN. 37
3.1 DeepCORE Framework 49
3.2 DeepCORE Architecture 50
3.3 Bi-directional LSTM Architecture (Cui *et al.*, 2018) 53
3.4 RNN Encoder-Decoder Architecture (Cho *et al.*, 2014b) 54
3.5 Attention Mechanism (Bahdanau *et al.*, 2014) 55
3.6 Comparison of Modeling Approach Between Expecto(a) and DeepCore(b) 57
3.7 Regression Error Rate Comparison of Various Models on E061 and E071 Cell Lines. The Dashed Red Line Represents the Error Rate of ExPecto. 61

Table Page

3.8 Binary Classification Performance of Various Models on E061 and E071 Cell Lines. The Dashed Blue Line Represents the F1 Score of DeepChrome. 62
3.9 Error Rate Comparison Between DeepCORE and ExPecto Across 25 Different Cell Lines. 64
3.10 Performance Comparison Between DeepCORE and DeepChrome Across 25 Different Cell Lines. 65
3.11 Cross-tissue Performance Comparison Between DeepCORE and DeepChrome Across 4 Cell Lines. F1 Score was Used as the Performance Metric. Black Lines Denote the Within Cell Line Accuracy. 66
3.12 Cross-tissue Performance Comparison Between DeepCORE and ExPecto Across 4 Cell Lines. MSE is Used as the Performance Metric. Black Lines Denote the Within Cell Line Error Rate. 67
3.13 CYFIP2 Gene in E007 Cell Line. (Top) Raw Histone Modification Read Count Signals. (Bottom) DeepCORE Attended Reegions. 68
3.14 CYFIP2 Gene in E058 Cell Line. (Top) Raw Histone Modification Read Count Signals. (Bottom) DeepCORE Attended Regions. 69
3.15 Density Plot of the Count of Attended Bins with Histone Signals Vs. Count of Attended Bins with No Histone Signals. 70

Table Page

3.16 Density Plot of the Frequency of Attended Bins with Known Promoters Across 25 Cell Lines. (a) Attended Bins with Promoters Vs. Random Attended Bins (B) Attended Bins with Promoters Vs. Attended Bins Without Promoters. 71

3.17 Density Plot of the Frequency of Attended Bins with Known Enhancers Across 25 Cell Lines. (a) Attended Bins with Enhancers Vs. Random Unattended Bins (B) Attended Bins with Enhancers Vs. Attended Bins Without Enhancers. 72

3.18 Mapping Attended Bins With Known Promoters and Enhancers for TRAK2 Gene in E006 Cell Line. 73

3.19 Mapping Attended Bins with Known Promoters and Enhancers for TMEM88 Gene in E004 Cell Line. 74

3.20 Fine Mapping of Genomic Regions in ARF5 Gene Across 3 Different Cell Lines. For Each Cell Line, the Top Portion Denotes Histone Signals and the Bottom Portion Represents the Attention. 75

3.21 Correlation Plot on PSMD8 Gene. 76

Chapter 1

INTRODUCTION

1.1 List of Publications

A portion of the research conducted as a part of this dissertation has already been published or in the press and the others are in preparation. The following are the list of publications that have either been used or influenced this dissertation for which I was fortunate to be one of the authors.

- [a] Liu, Li, Maxwell D. Sanderford, Ravi Patel, **Pramod Chandrashekar**, Greg Gibson, and Sudhir Kumar. "Biological relevance of computationally predicted pathogenicity of noncoding variants." Nature communications 10, no. 1 (2019): 1-11.

- **Chandrashekar, Pramod**, Navid Ahmadinejad, Junwen Wang, Aleksandar Sekulic, Jan B. Egan, Yan W. Asmann, Sudhir Kumar, Carlo Maley, and Li Liu. "Somatic selection distinguishes oncogenes and tumor suppressor genes." Bioinformatics 36, no. 6 (2020): 1712-1717.

In addition to these, I have worked on several other projects with the following publications:

- Sarker, Abeed, **Pramod Chandrashekar**, Arjun Magge, Haitao Cai, Ari Klein, and Graciela Gonzalez. "Discovering cohorts of pregnant women from social media for safety surveillance and analysis." Journal of medical Internet research 19, no. 10 (2017): e361.

[a] Included in Section 1.4.2

- **Chandrashekar, Pramod Bharadwaj**, Kathleen Montone, Michael D. Feldman, and Graciela Gonzalez-Heranndez. "Automated comparison of pathology reports for on-the-job assessment of residents." AMIA Summits on Translational Science Proceedings 2019 (2019): 495.

1.2 Overview of Gene Regulatory Mechanism

Gene Regulation is a set of mechanisms that induce or repress the expression level of genes. These mechanisms involve transcription, translation, and post-translational modifications (Curtis *et al.*, 1995; Appella and Anderson, 2001). Transcriptional regulation specifically refers to the transcriptional process during which DNA is copied into messenger RNA (mRNA). In Eukaryotes, Two regulatory components are involved:

- Cis-regulatory elements
- Trans-regulatory elements

1.2.1 Cis-regulatory Elements

Cis-regulatory elements are coding or noncoding sequences that can be bound by various proteins (i.e., transcription factors, TFs) to modulate transcription (Wittkopp and Kalay, 2012). These elements include promoters, enhancers, silencers, and insulators, and are divided into two categories based on the distance to the transcription start site (TSS) of a gene (**Figure 1.1**). Proximal regulatory elements are very close to TSSs, usually within 1 Kilobase-pairs (Kbps), and are mainly comprised of promoters. Distal regulatory elements are located as far as 1 Megabasepairs (Mbps) from TSSs and include enhancers, insulators, silencers, and distal promoters.

Proximal Regulatory elements are very close to the Transcription Start Site (TSS)

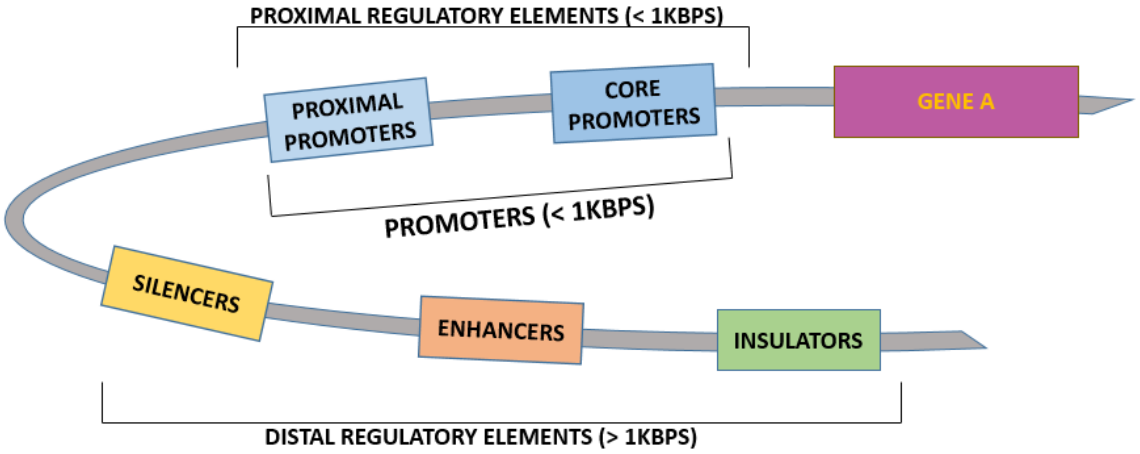


Figure 1.1: Cis-regulatory Elements (Modified from Maston *et al.* (2006))

of the gene, usually, <1 Kilobase-pairs (Kbps) of the gene and are mainly comprised of promoters. Distal Regulatory elements are regions that can be as far as 1 Megabase-pairs (Mbps) from the TSS and include enhancers, insulators, and silencers.

Promoters

Promoters are DNA regions where transcription initiates. Promoters are comprised of two components: core promoters and proximal promoters (Riethoven, 2010). Core promoters contain TSSs and are docking sites of RNA polymerases. Various core promoters have been discovered with the most famous one being the TATA box. Core promoters cannot mediate transcription by itself and need proximal regions to bind to various TFs to function properly. A popular set of sequences that act as proximal promoters is the CAT box. Recently distal promoters have been discovered at several Kbps away from TSSs (Chen *et al.*, 2010).

Insulators

Insulators act as barriers such that the target genes are not transcribed by regulatory elements controlling other nearby genes (Maston *et al.*, 2006). They define gene transcription regions and hence are called boundary elements. Insulators function in two ways. Firstly, they restrict transcription by blocking enhancer-promoter interaction (Zhao and Dean, 2004). Secondly, they block the spread of the repressive chromatin into transcriptional regions (West *et al.*, 2002). The CTCF binding factor is the only known TF that binds to insulators in mammals (Chatterjee and Ahituv, 2017).

Enhancers

Enhancers are usually far (up to hundreds of Kbps) from TSSs of their target gene (Krivega and Dean, 2012; Blackwood and Kadonaga, 1998). They play a major role in spatial, temporal, and tissue-specific gene expression regulation. These enhancers contain a cluster of TFBSs that work coordinately to regulate transcription. They act as modules that are independent of orientation and distance from TSSs (Khoury and Gruss, 1983; Narlikar and Ovcharenko, 2009; Maston *et al.*, 2006) and can be in the upstream, downstream, or intron regions of their target genes.

Enhancers are proposed to work under two mechanisms. First, it can change the chromatin structure which facilitates the binding of various TFs that look along the DNA scanning for target promoters (Narlikar and Ovcharenko, 2009; Blackwood and Kadonaga, 1998). Second, it interacts with a promoter directly through 'DNA looping' where the DNA between an enhancer and a promoter is looped out, which enables protein-protein interactions between TFs for gene activation (Narlikar and Ovcharenko, 2009; Carey *et al.*, 2009).

Enhancers control tissue-specificity of gene expression because a unique set of

enhancers is recruited in different tissue and cell types (Chatterjee and Ahituv, 2017). Therefore, they are integral components of tissue-specific genetic circuits that underlie complex developmental events in higher eukaryotes. A key mechanism that enables tissue-specific functions of an enhancer is through tissue-specific expression of TFs that bind to it. Alternatively, changes of chromatin structure may block the access of TFs to an enhancer even if the TFs are highly expressed.

Silencers

These regulatory elements have negative effects on gene regulation. The main function of silencers is to repress the activities of the gene (Maston *et al.*, 2006). They compete with enhancers for the TFs to bind. They can be located close to the promoters or far away like the enhancers. There are two known mechanisms by which they repress gene activity: a) by blocking TF binding and/or competing with enhancers for the same site, b) by injecting repressive chromatin structures preventing promoter activator communication (Maston *et al.*, 2006; Hoppe and Francone, 1998; Chen and Widom, 2005).

1.2.2 Trans-regulatory Elements

Trans-regulatory elements are genes that modulate the expression of other genes. They include genes coding for TFs, genes involved with chromatin modification, etc. These elements can initiate or repress transcription.

Transcription Factors

TFs are specific proteins that bind to DNA sequences (i.e., TFBSs) involved in transcriptional regulation (Hobert, 2008). A single TFBS can bind to multiple TFs and a TF can bind to multiple sites. A regulatory complex consists of multiple TFs interact-

ing with each other, which creates a 3-D structure bringing distant genetic elements into close vicinity (**Figure 1.2**).

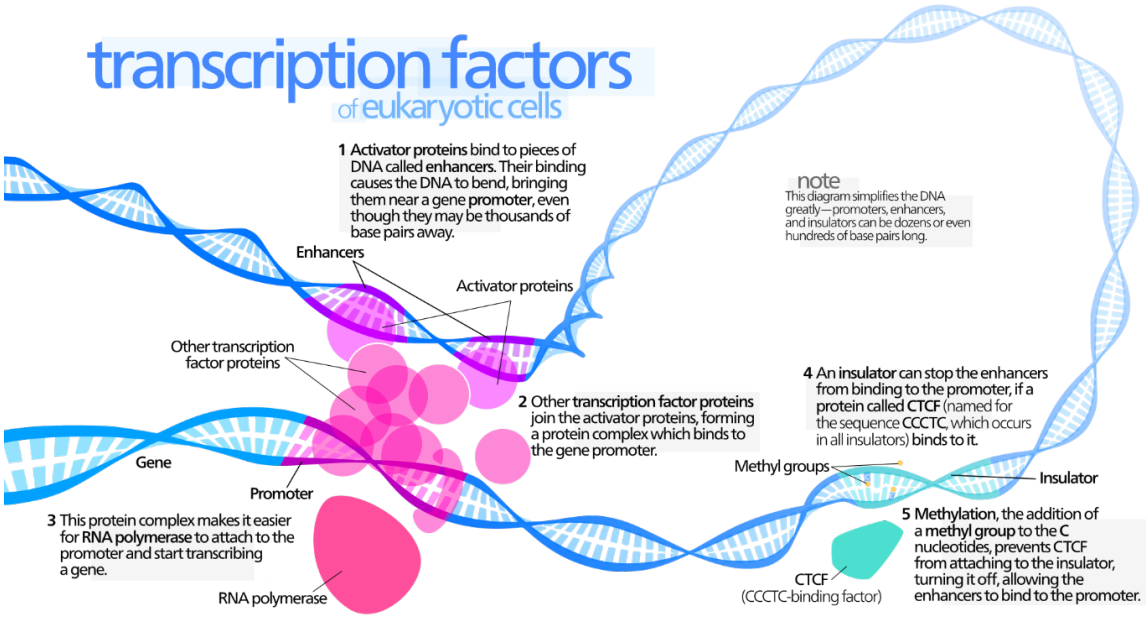


Figure 1.2: Transcription Factors Binding to the Regulatory Regions Initiating Transcription (Blackwood and Kadonaga, 1998)

For example, in a regulatory complex, TFs concurrently bind to the DNA promoter near the TSS and an enhancer further upstream of a gene. Depending on the expression levels of the TFs in this complex in a specific tissue, transcription of the target gene may be stimulated or repressed (Chen and Widom, 2005).

Histone Modifications

Chromatin modification plays a major role in gene regulation. Chromatin contains nucleosomes where the DNA is packaged with histone proteins. The core histone proteins are *H2A, H2B, H3, and H4*. These nucleosomes are governed by chromatin-modifying factors and co-factors that are histone-modifying enzymes and other supportive molecules. These (co)-factors modify chromatin structure and subsequently affects the binding of TFs to the regulatory elements. Various chromatin modifi-

cation types are shown in **Table 1.1**. When histone modifications (HMs) alter the chromatin structure of regulatory elements, the accessibility of the affected regulatory element to TFs may change, which subsequently alter the transcription level of the target gene.

Chromatin Modifications	**Residues Modified**	**Functions Regulated**
Acetylation	**K**-ac	Transcription, Repair, Replication, Condensation
Methylation (lysines)	**K**-me1, **K**-me2, **K**-me3	Transcription, Repair
Methylation (arginines)	**R**-me1, **R**-me2a, **R**-me2b	Transcription
Phosphorylation	**S**-ph, **T**-ph	Transcription, Repair, Condensation
Ubiquitylation	**K**-ub	Transcription, Repair
Sumoylation	**K**-su	Transcription
ADP ribosylation	**E**-ar	Transcription
Deimination	**E > Cit**	Transcription
Proline Isomerization	**P-Cis > P-trans**	Transcription

Table 1.1: Different Classes of Histone Modifications (Kouzarides, 2007)

Fig. 1.3 represents the various combination of histone modifications in the promoter and enhancer regions. Active enhancers are characterized by modifications involving H3K27ac which are associated with promoters baring H3K27ac and H3K4me3. This combination leads to gene activation. Poised enhancers have H3K27me3 instead of H3K27ac and H3K4me1. They are associated with promoters having H3K4me3, H3K27me3, and H3K9m3. This leads to gene deactivation. Similarly, repressed enhancers and promoters both have H3K27me3 which also results in gene deactivation.

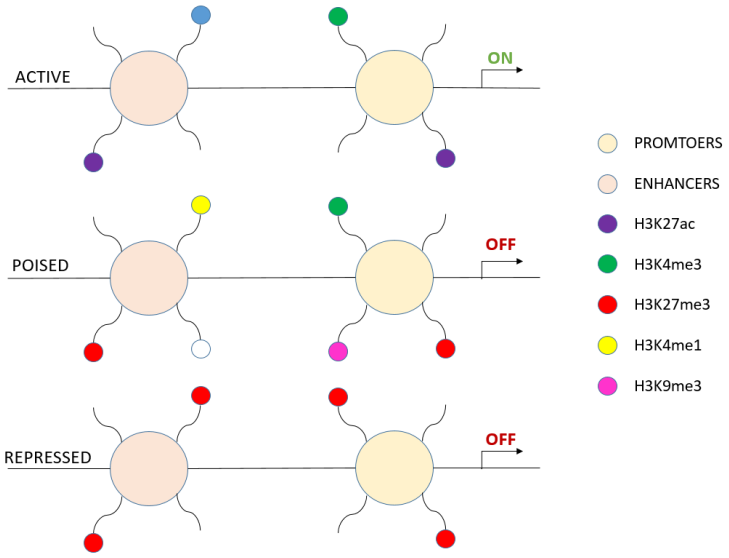


Figure 1.3: Histone Modifications Affecting Gene Regulation. (Top) Actively Transcribed Genes Are Marked by H3k27ac and H3k4me3. (Middle) This Represents a Poised/Bivalent State Where We Have Both Activators (H3k4me1) in the Enhancer Region and Repressors(H3k9me3, H3k7me3) in the Promoters Leading to Repression of Genes. (Bottom) Presence of H3k27me3 in Both Enhancers and Promoters Always Leads to Gene Repression.

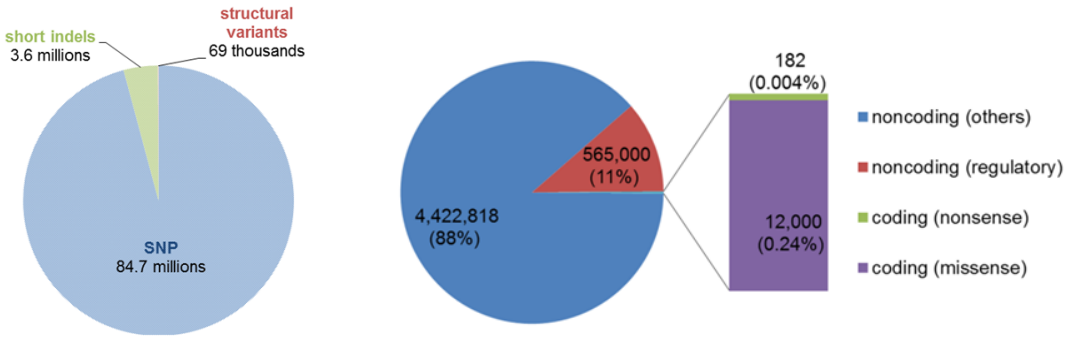
1.3 Genotype-Phenotype Association

1.3.1 Genetic Mutations

Genetic variants are permanent changes in DNA sequences that underlie phenotypic differences between and within species (Spielman *et al.*, 2007; Frazer *et al.*, 2009). Two types of mutations exist:

- • Germline Variants - They are present in the reproductive cells of the parents. A person inherits these variants. All cells in the person carry these germline variants.

- Somatic Mutations – They are acquired after birth and accumulate throughout a person's lifetime. Different cells in a person may carry different somatic mutations.



(a) Pie chart showing the various kinds of genetic mutations.

(b) Pie chart showing the different regions where the genetic mutations can occur.

Figure 1.4: Types of Genetic Mutations

Germline variants can be classified into different categories based on structural impact on DNA sequences. The vast majority of these variants are single nucleotide variants (SNVs) that affect a single base pair (**Figure 1.4a**). They can be classified based on functional impact (**Table 1.2**) and can also be classified in general based on the its types (**Table 1.3**). A huge portion of these genetic variants is present in the noncoding regions with unknown functions (**Figure 1.4b**).

Region	**Functional Impact**
Coding	Synonymous, Nonsynonymous, silent, truncating
Non-coding	Regulatory, non-regulatory

Table 1.2: Types of Mutations Based on Functional Impact

If a genetic variant cause significant functional changes, it may lead to phenotypic differences including many related to human health. Traditional research of genetic variants that focuses on protein-coding regions has led to the discovery of thousands of disease-associated variants (Lek *et al.*, 2016; MacArthur *et al.*, 2012). Recently the functional roles of noncoding variants in human diseases and quantitative traits have

Single point mutation	Insertion	An extra base is added to the existing sequence
	Deletion	A base is removed from the sequence
	Substitution	A base is replaced by another base
	Inversion	A region of chromosome is replaced in reverse order
	Deletion	A portion of chromosome is deleted
Block Mutation	Duplication	A region of chromosome is repeated only once
	Translocation	Different segments of chromosome are exchanged
Copy Number Variation	Repeats	A segment of chromosome is repeated multiple times

Table 1.3: Classes of DNA Mutations and Their Impact

been increasingly recognized (Maston *et al.*, 2006; Chatterjee and Ahituv, 2017). For example, a mutation in the promoter region of the GpIbB gene leads to Bernard-Soulier syndrome. Many cancer subtypes are known to be caused by changes in regulatory elements. For example, a mutation in the cMYC gene is associated with T-cell leukemia.

1.3.2 GWAS and eQTL Analysis

Genome-wide association studies (GWAS) investigate genotype-phenotype associations using genotyping microarray and whole-genome sequencing (WGS) techniques. A typical GWAS profiles millions of germline genetic variants in many individuals with different phenotypes and determines the association of these variants with heritable diseases (Visscher, 2008; Nicolae *et al.*, 2010). Although GWAS have discovered many genetic loci associated with human diseases and quantitative traits, they have two main limitations. Firstly, genetic variants have a wide range of population frequencies and effect sizes. However, most of the disease-associated variants (DAVs) identified by GWAS have high frequencies and small effect sizes (**Figure 1.5**)

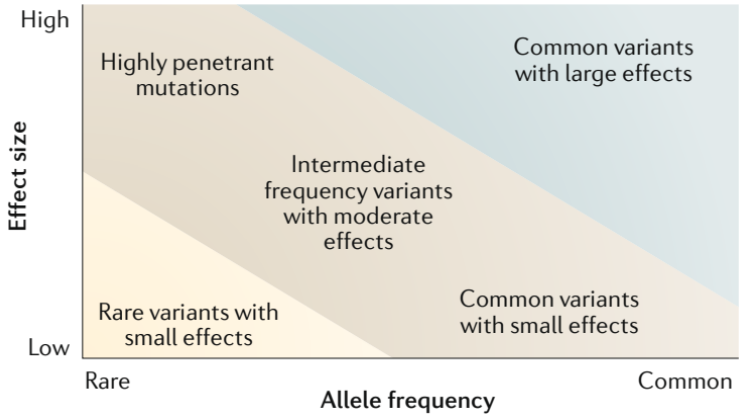


Figure 1.5: Spectrum of Genetic Variants.

GWAS have notoriously low power to discover rare DAVs (Tam *et al.*, 2019). Secondly, a genomic locus usually harbors hundreds to thousands of SNVs with strong linkage. While GWAS may point to the disease-associated locus, it cannot pinpoint the causal variants. All DAVs discovered to date collectively account for only a small portion of heritability. For example, Chron's disease is associated with close to 43 different loci and yet only 20% of phenotype variance can be explained (Barrett *et al.*, 2008). Another example is height, a very common trait. It is associated with

40 different loci and yet only 5% of the variance is explained (Visscher, 2008). The missing heritability calls for a new research focused to discover novel DAVs, fine mapping of causal variants, and epistasis of genetic variants (Eichler *et al.*, 2010; **?**; **?**).

Expression quantitative trait loci (eQTL) mapping is another approach similar to GWAS, except that it studies a phenotype at the molecular level, i.e., genetic variants associated with gene expression levels (Gilad *et al.*, 2008; Li *et al.*, 2018). They have proved to be a key step in investigating regulatory pathways and have identified numerous eQTLs that modulate the expression of genes associated with various human diseases like diabetes. Despite its huge success, it suffers the same drawbacks as GWAS. Furthermore, because GWAS and eQTL studies are based on associations, the identified DAVs need to be further validated that can be extremely time consuming and cost-inefficient. Computational methods that accurately prioritize candidate DAVs will significantly accelerate the study of genotype-phenotype associations.

1.4 Computational Methods for Genotype-Phenotype Studies

1.4.1 Computational Methods to Discover DAVs and eQTL Variants

Computational methods are an indispensable component enabling GWAS and eQTL studies by providing efficient and effective analyses to discover DAVs and prioritize causal variants for further validation. These approaches involve sophisticated statistical modeling and machine learning techniques. For example, Sun *et al.* (2007) used regression and random forest to identify and predict genome wide SNP effect on rheumatoid arthritis. BEAM (Zhang and Liu, 2007) used Bayesian partitioning and markov chain Monte Carlo approach to associated multiple genotypic markers with age-related macular disease. PLINK (Purcell *et al.*, 2007) was an all-in-one soft-

ware developed to manage large volumes of data and also provided various statistical association tools for whole genome association studies (WGAS).

Most existing methods assume independence of DAVs, thus examine the association of an individual genetic variant with the phenotype. However, because most human diseases are polygenic and most genes are regulated by multiple regulatory elements, such an assumption is an oversimplification. To overcome this limitation, BOOST (Wan *et al.*, 2010) used Boolean representation of genome-wide SNPs to study all possible pairwise interactions. Ferreira and Marchini (2011) modeled the evidence of SNP associations by averaging over several developed models accounting for SNP interactions in the risk loci. Recently, as it has become well recognized that epistasis is ubiquitous, interactions involving more than two variants may have significant contributions to the missing heritability. Therefore, new computational methods that examine the combinatorial effects of multiple genetic variants will significantly advance the genotype-phenotype research and accelerate the discovery of novel DAVs.

1.4.2 Computational Methods for Predicting the Functional Impact of Genetic Variants

Once DAV or eQTL loci are discovered, the next big question is which variants in these loci are causal? For a variant to be causal, it must have significant functional impact. There are several methods that predict the functional impact or pathogenicity of a genetic variant and can be grouped into two categories based on if the variant is located in a coding region or a noncoding region. Computational approaches to predicting functional coding variants are well developed. Methods for finding pathogenic non-synonymous SNVs (nsSNVs) include SIFT (Ng and Henikoff (2003), SVM based tool), PolyPhen (Adzhubei *et al.* (2013), a naïve bayes approach), EvoD (Kumar *et al.* (2012), an evolution-based regression model), and REVEL (Ioannidis *et al.*

(2016), uses an ensemble technique to predict the pathogenecity of missense mutations), among several others. For synonymous SNVs (sSNVs), SilVA (Buske *et al.*, 2013) uses random forest with conservation, sequence, and RNA features to predict the harmfulness of synonymous SNVs (sSNVs). TraP (Gelfman *et al.*, 2017) predicts the sSNVs damage to the gene transcript. RegSNPs (Zhang and Liu, 2007) predicts the impact of sSNVs on protein function and mRNA splicing.

Compared to coding variants, predicting pathogenic noncoding SNVs (ncSNVs) has been a long-standing challenge. Although several computational tools have been developed, our systematic evaluation of the biological relevance of these predictions have pointed to a significant knowledge gap (Liu *et al.*, 2019). We summarize in Table 1.4 our survey of six current tools, namely CADD (Kircher *et al.*, 2014), CATO (Maurano *et al.*, 2015), DeepSEA (Zhou and Troyanskaya, 2015), EIGEN (Ionita-Laza *et al.*, 2016), GWAVA (Ritchie *et al.*, 2014), and LINSIGHT (Huang *et al.*, 2017).

Regardless of the different theoretical and empirical aspects of each method, these tools invariably claim excellent discrimination of pathogenic from non-pathogenic variants (reported AUROC values up to 0.97, **Table 1.4**), which has encouraged their application in over a thousand studies of the genetic basis of biomedical phenotypes.

Unfortunately, the concordance of pathogenicity predictions made by current tools is low (Wang and Long, 2016) and in vitro evidence is frequently at odds with in-silico assessments (Kircher *et al.*, 2014). These observations suggest that the accuracies reported during the development of predictive methods may not translate into high success rates when applied to fundamental empirical biological research questions. To investigate the underlying reasons of the large discrepancy, we designed two biologically relevant tasks that point toward needed areas for improvement.

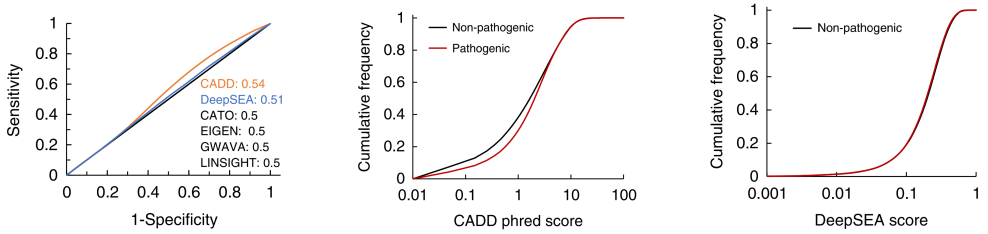
Table 1.4: Properties of Predictive Models for Six Tools

Method	**Predictors**	**Modeling Approaches**	**Performance (AUROC)**
CADD	Evolutionary parameters, ENCODE summaries, functional annotations, population frequencies	Support Vector Machines	0.92
CATO	Cell type- and tissue-specific assays, evolutionary parameters, functional annotations	Logistic regression	NA
DeepSEA	Sequences, evolutionary parameters	Deep learning, Logistic regression	0.85
EIGEN	Evolutionary parameters, ENCODE summaries, population frequencies	Unsupervised learning	0.79
GWAVA	Evolutionary parameters, ENCODE summaries, population frequencies	Random Forests	0.97
LINSIGHT	Evolutionary parameters, ENCODE summaries, functional annotations	Generalized linear model	0.96

Task 1: Discriminating pathogenicity of alternative alleles

Given a genomic position with segregating alleles in a population, except in rare instances of balancing selection, one allele will usually be more deleterious than the other, presumably contributing to pathogenesis (Cai *et al.*, 2009; Josephs *et al.*, 2015;

Zeng *et al.*, 2018). At any position, apart from the reference allele, we need to know which of the other three bases are pathogenic, because contrasts of their allele frequencies in different populations and/or relative to the ancestral state can help to elicit the direction of natural selection and shed light on the genetic basis of health disparities (Corona *et al.*, 2010; Raj *et al.*, 2013; Polimanti and Gelernter, 2017). Our expectation was that a pathogenic allele would receive a significantly higher impact score (as defined for each of the six tested methods) than a non-pathogenic allele at the same position. Instead, we found that these methods were unsuccessful at this task. In fact, four of them (LINSIGHT, EIGEN, GWAVA, and CATO) reported identical scores for all alternative alleles at every position as they were not designed for allelic contrasts (**Figure 1.6a**).



(a) (b) (c)

Figure 1.6: Performance of Tested Methods on Detecting Pathogenic Variants in Position-matched Ncsnvs. (a) ROC Curves with AUROC Values Displayed for Each Method. (b) Cumulative Distribution of CADD Scores. (c) Cumulative Distribution of DeepSEA Scores.

Two methods (CADD and DeepSEA) produced Different scores for pathogenic and non-pathogenic alleles at a position, but the AUROC was only slightly higher than 0.5 (AUROC = 0.54 and 0.51, DeLong test p-value = 10^{-11} and 10^{-13}, respectively). Despite the statistical significance, the small effect sizes implied that the biological usefulness of these methods for task #1 is extremely limited, which was further confirmed by the statistically significant yet largely overlapping distributions of impact scores for pathogenic and non-pathogenic alleles (Wilcoxon signed-rank test

p-value = 0, **Figures 1.6b and 1.6c**). Therefore, it is apparent that none of the six methods are suitable for distinguishing among the alleles segregating at a site with respect to deleteriousness.

Task 2: Resolving the identity of causal alleles

We examined if the six prediction methods can distinguish pathogenic from non-pathogenic positions within credible intervals defined by physical distance or linkage disequilibrium. Between each pair of a DAV and a nearby CPP, a successful method should assign a higher impact score to the DAV. However, we were unable to reject the null hypothesis for CATO when the distance between a DAV and a CPP is less than 1 kbps, for EIGEN when distance < 100 bps, for GWAVA when distance < 50 bps, and for CADD, DeepSEA, and LINSIGHT when distance < 10 bps (paired one-sided t-test p-value>0.05). Even when the statistical tests reached the significance threshold, the effect size represented by the score difference between a pair of DAVs and CPPs was usually very small (**Figure 1.7a**). For example, the DeepSEA score of a DAV, on average, was only 0.02 higher than its nearby CPP, while the full range of the DeepSEA score is between 0 and 1. Among all tested methods, CADD and LINSIGHT had relatively large positive components. However, even for these two methods, less than a quarter of the evaluations are greater than a standard deviation unit.

As the physical distance increases, the differences between the impact scores for DAVs and CPPs generally become more evident, i.e., pathogenic variants can be distinguished from nonpathogenic variants when they are located further apart (**Figure 1.7a**). These results are likely explained by the high correlation between the impact scores of DAVs and matching adjacent CPPs, which decreases with increasing physical distance (**Figure 1.7b**). For different tests, this may reflect contributions of

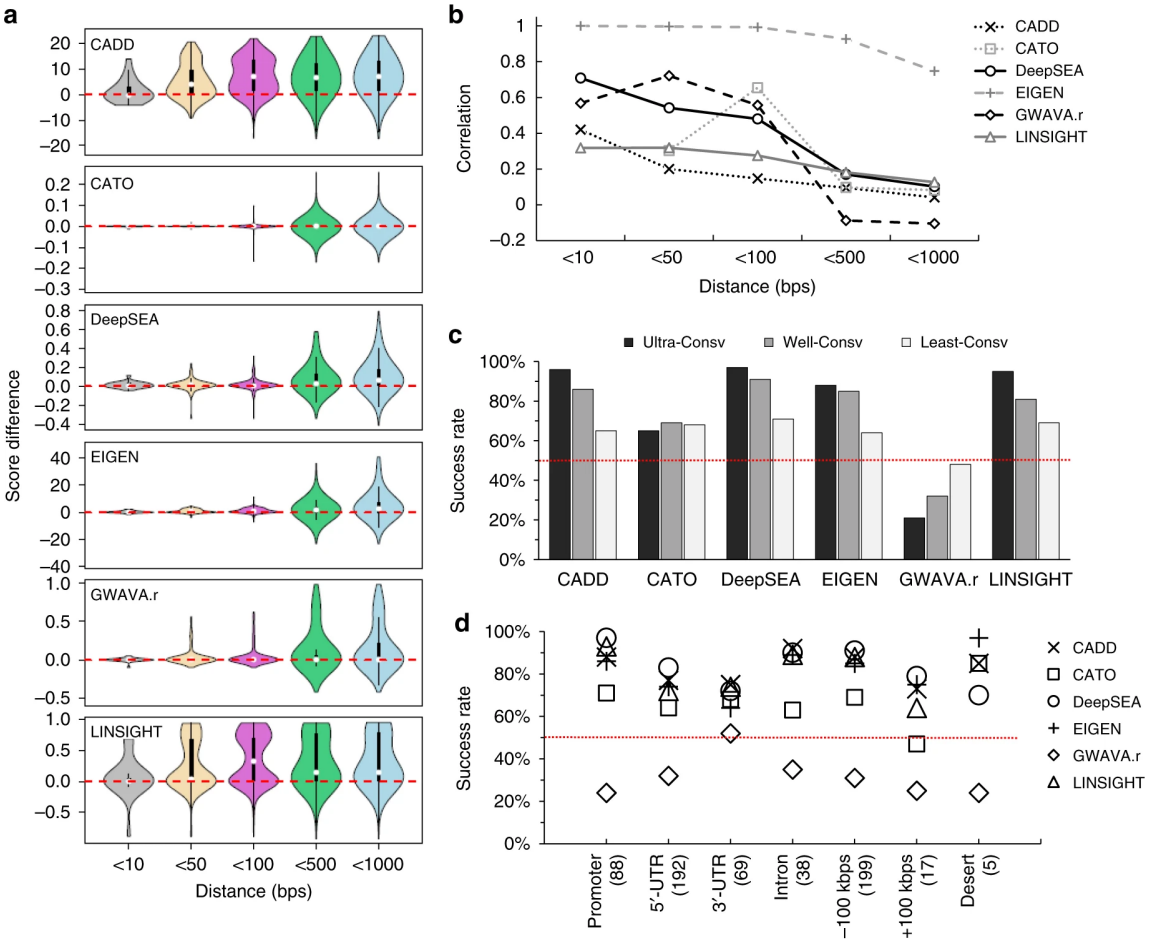


Figure 1.7: Performance of Tested Methods on Region-matched Pathogenic and Non-pathogenic ncSNVs. (a) Violin Plots Show Distributions of Impact Score Difference Between Nearby Pathogenic and Non-pathogenic ncSNVs. (b) Correlation of Impact Scores for ncSNVs Located Within given Genomic Distances. (c) Success Rates of Ranking a Pathogenic ncSNVs Higher than a Non-pathogenic ncSNVs Located Within 1000bps Vicinity. (d) Success Rates Stratified by the Genomic Context of Pathogenic ncSNVs.

evolutionary history of the haplotype block yielding similar conservation patterns, or the likelihood that closely linked alleles all lie within an extended regulatory stretch of chromatin with similar epigenetic marks.

We also stratified the data by using the evolutionary conservation of the position where a DAV is found: ultra-, well-, and least-conserved categories (absolute evolutionary rate $= 0$, < 2, and ≥ 2 substitutions/site/billion years, respectively). As

reported previously, the maximum success rate of ranking a pathogenic ncSNV higher than an adjacent ncSNV was achieved when the variant affected an ultra-conserved position, which is up to twice as good as that at the least-conserved positions (**Figure 1.7c**).

Among different types of regulatory elements, the prediction methods work best in identifying pathogenic ncSNVs disrupting a promoter (**Figure 1.7d**). This may be because promoter regions are among the best studied regulatory elements in the genome, and most of the tools provide quite good discrimination for up to 97% of true DAVs in promoters. The improved performance may be attributed to the conservation of promoters (Xie *et al.*, 2005; Villar *et al.*, 2015) reflecting the strong evolutionary component of the impact scores.

Our biologically oriented evaluations indicate an urgent need for computational methods that can accurately fine map functional genetic variants in noncoding regions.

1.4.3 Machine Learning and Deep Learning in Bioinformatics

Supervised and unsupervised learning has been in use for a very long time. Deep learning, a relatively newer advancement, has proven to be more efficient than the traditional approaches as it allows for the stacking of multiple layers and can learn high dimensional complex data. Deep Learning has found great success in computer vision for a wide range of tasks like motion tracking (Doulamis and Voulodimos, 2016; Doulamis, 2018), human activity recognition (Toshev and Szegedy, 2014; Lin *et al.*, 2016; Cao and Nevatia, 2016), autonomous driving (Chen *et al.*, 2015a; Uçar *et al.*, 2017; Prabhakar *et al.*, 2017; Boloor *et al.*, 2019), and object recognition (Sun *et al.*, 2015; Ren *et al.*, 2015; Redmon *et al.*, 2016). They have also found great success in natural language processing for various tasks like machine translation, named entity

recognition, question answering systems, among other tasks (Luo *et al.*, 2015; Chiu and Nichols, 2016; Dong *et al.*, 2015; Bahdanau *et al.*, 2014).

Machine learning and deep learning have been extensively used in the field of bioinformatics. As assessing the gene and protein function is a tedious and time consuming task in wet laboratories, machine learning is often a useful approach for predicting the functionality of gene and proteins in terms of disease relationship. Several methods exist in this space. Barutcuoglu *et al.* (2006) used the hierarchical structure of the gene functional class to build a Bayes net and support vector machine classifier to classify gene functions. As similar approach was used by Sokolov and Ben-Hur (2010) to identify protein function by modeling the structure of the proteins from Gene Ontology into kernel space. Kulmanov *et al.* (2018) used a deep convolution neural network (CNN) with protein sequences as input to predict the protein function. A similar but slightly different area is the identification of disease related genes. In this space, Li *et al.* (2019) extracted potential harmful genes associated with Triple Negative Breast Cancer (TNBC) by extracting a handful of genes using protein-protein interaction network and giving them as features to support vector classifier. Xuan *et al.* (2019) took a different approach where their goal was to identify long non-coding RNA genes (lncRNA) responsible for various diseases. They integrated disease similarities, lncRNA gens, and several interactions between them and modeled them as input to an attention-based convolution neural network model to identify genes responsible for 402 different diseases.

Text mining is another area in bioinformatics where the task is to look for associations (like protein-protein interaction, gene-disease association, etc) in biomedical literature for which various natural language processing techniques are used (Zhang *et al.*, 2006; Hu *et al.*, 2008; Wei *et al.*, 2013; Leaman *et al.*, 2015; Xu *et al.*, 2018; Lamúrias, 2019). Biomedical image analysis is another area which requires machine

learning and deep learning as it can handle huge volumes of high-dimensional data and is able to accurately analyze these images. Advancement in the technology has led to development of various models for identifying and learning various characteristics and features (Wu *et al.*, 2013; Liao *et al.*, 2013; Guo *et al.*, 2015), image segmentation for removal of unwanted regions (Prastawa *et al.*, 2005; Long *et al.*, 2015; Kleesiek *et al.*, 2016), detection of abnormal portions of the image to aid clinicians (Krizhevsky *et al.*, 2017; Ciompi *et al.*, 2015; Setio *et al.*, 2016; Gao *et al.*, 2018), and diagnosis of the disease based on the image (Kim *et al.*, 2016; Jeyaraj and Nadar, 2019; Horie *et al.*, 2019).

Deep learning is also used for various other tasks like predicting regulatory elements involving promoters, enhancers, and interactions between them. Umarov and Solovyev (2017) used gene sequences to train a CNN model to predict if they were promoters for five different species. Eser and Churchman (2016) built FIDDLE which passes multiple data sources (TF chip-seq, RNA-seq, and DNA sequences) into CNN architecture to predict the TSS of the gene in yeast organisms. DeePromoter (Oubounyt *et al.*, 2019) used an advanced deep learning model where they coupled CNN architecture with Long Short Term Memory (LSTM) network with gene sequences as input to predict promoter site in humans and mouse. SPEID (Singh *et al.*, 2019) uses a very similar architecture (CNN-LSTM) to predict the promoter-enhancer interaction using genetic sequences alone. Another area is TFBS prediction. DeepBind (Alipanahi *et al.*, 2015) uses CNN architecture to model the association between sequences and its binding to various TFs and then extracted the features to visualize the Motifs identified across these sequences. TFImpute (Qin and Feng, 2017) uses the same CNN architecture with additional gating mechanisms for the task of TF binding predictions. Comprehensive reviews of machine learning and deep learning in different areas of bioinformatics is found in (Holzinger *et al.*, 2014; Park

and Kellis, 2015; Angermueller *et al.*, 2016; Jones *et al.*, 2017; Shen *et al.*, 2017; Litjens *et al.*, 2017; Fourcade and Khonsari, 2019).

Attention-based deep learning has been widely used ever since it was first used in neural machine translation task (Bahdanau *et al.*, 2014). They are mostly used for extracting useful information from clinical text. For example, Peng *et al.* (2019) used self-attention based deep learning model to extract lesion attribute information from radiology reports, Chen (2019) built an attention-based BiLSTM model for assertion detection in clinical notes, and Tang *et al.* (2019) implemented attention mechanism to classify clinical progress notes into six different clinical categories, among others. They are also used in imaging informatics. Tomita *et al.* (2019) built attention-based deep learning model to identify cancerous esophagus tissue in histological images. They are used for image segmentation tasks (Nie *et al.*, 2018; Chen *et al.*, 2018). However, we have not seen applications of attention-based deep learning to predicting regulatory elements or pathogenic genetic variants, which leaves a promising topic for us to investigate.

1.5 Conclusion

As every gene can have multiple regulatory regions, there can be several SNPs across these regions that can contribute together in gene regulation or disease enhancement. As the existing methods do not consider the combinatorial effect of SNPs, new computational methods for exploring the combinatorial effects of SNPs on gene regulation are needed.

Chapter 2

TREEMAP: A STRUCTURED APPROACH TO FINE MAPPING OF EQTL VARIANTS

Scans for expression quantitative trait loci (eQTL) aim to discover genetic variants associated with variation in transcript abundance among individuals. Genome-wide scanning of eQTL involves genomic and transcriptomic profiling of a large number of samples, followed by statistical and experimental analyses of polymorphic sites to discover causal variants (Gaffney *et al.*, 2012). Due to linkage disequilibrium (LD), typically multiple genetic variants at a locus show highly significant statistical scores, although only some of these are causal. Expression-associated variants (eVars) are usually aggregated into a credible set that includes a lead variant with the strongest association signal and other linked variants. However, a lead eVar is not necessarily responsible for transcriptional regulation, but tags causal variants instead (Schaid *et al.*, 2018; Van de Bunt *et al.*, 2015). Furthermore, in genes with multiple cis-acting eQTL, the correspondence between lead eVars and causal variants diminishes quickly due to the superimposed effects of adjacent loci (Zaykin and Zhivotovsky, 2005; Zeng *et al.*, 2019).

2.1 Background

To better resolve causal variants, recent fine-mapping efforts have gone beyond the conventional single-site assumption and evaluated multisite effects. Because an exhaustive search for an unknown number of causal variants in a wide genomic region is computationally prohibitive, several strategies have been employed to ease the computational burden. Stepwise conditional regression is a greedy algorithm that

repeatedly tests individual sites and returns lead eVARs with the best marginal test statistics at each iteration (Yang *et al.*, 2013). This algorithm is computationally efficient although the solution is highly susceptible to local optima. CaVEMaN (Brown *et al.*, 2017), CAVIARBF (Chen *et al.*, 2015b), FINEMAP (Benner *et al.*, 2016), and PAINTOR (Kichaev *et al.*, 2014) apply sophisticated resampling and search strategies to explore additional causal configurations among candidate variants. However, these methods require prior knowledge of potential causal variants that are often derived from single-variant or stepwise association tests. The adaptive DAP method takes a tiered strategy. It first scans a genomic region for independent eQTL and then conducts an exhaustive search within each locus (Wen *et al.*, 2016). Although this method does not impose constraints on the number of causal variants, attempts at finding more than four causal variants are still computationally intensive (Zeng *et al.*, 2017). Given that most human genes have multiple cis-acting eQTL (Ulirsch *et al.*, 2019; van Arensbergen *et al.*, 2019) and independent studies have reported that credible intervals generally contain one hundred or more eVars per gene (Bhalala *et al.*, 2018; Kim *et al.*, 2014; Strunz *et al.*, 2018), fine-mapping algorithms capable of identifying an arbitrary number of eQTL, prioritizing multiple eVars at a locus, and performing at high computational efficiency will improve genome-wide discovery of regulatory variants.

While LD between eVars adds to the complexity of eQTL fine-mapping, it also provides a convenient structure with which large genome regions can be dissected into multiple relatively independent segments that are then amenable to association testing. Intuitively, one can partition variants into small blocks based on a specific cutoff value, although the most appropriate cutoff is unknown a priori. Furthermore, a clear boundary is often hard to find at loci with a non-monotonic LD structure where highly correlated variants are interspersed with uncorrelated variants (Daly *et al.*,

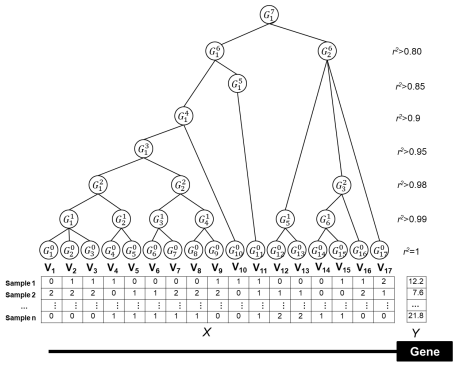
2001). The Tree Scanning method addresses this issue by organizing genomic regions into a hierarchical tree to study phenotypic associations (Templeton *et al.*, 2005). However, because this method uses haplotypes as the genomic unit, it lacks base-pair resolution and is unsuited to fine-mapping tasks. Tree-guided lasso (Yuan *et al.*, 2011) offers an intuitive solution, in which the selection of groups of variants or individual variants is conducted in a hierarchical framework defined by LD structure. This machine-learning method is also highly efficient for genome-scale analysis. However, it does not provide statistical confidence on the selected features required for biological and clinical applications.

To address these deficiencies, we designed a nested model that first employs the tree-guided lasso algorithm to scan a large genomic region for candidate loci and candidate variants within a locus, then apply statistical inference to derive credible sets of putative causal variants. We tested this new method, named TreeMap, via rigorous simulations. We show that TreeMap has significantly higher accuracy and faster computation than existing methods under various scenarios, especially for genes with multiple cis-acting eQTL under weak to medium LD. Applications of TreeMap to GTEx data of brain hippocampus samples and transverse colon samples revealed abundant distal regulatory variants located in up to 2 Mbps away from gene bodies.

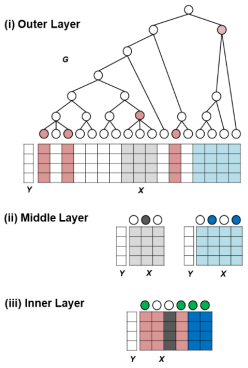
2.2 Materials and Methods

2.2.1 Data Structure

Given samples, each genotyped at biallelic positions in the upstream region of a target gene, a feature matrix contains genotype data with rows corresponding to samples and columns corresponding to genetic variants. A response vector contains expression level of the target gene in samples. To represent the LD structure of the



(a) Data Structure



(b) Nested Design

Figure 2.1: The TreeMap Method

variants, we compute the squared correlation coefficient (r^2) between pairs of variants. We define six (r^2) cutoffs ($>$ 0.999, 0.98, 0.95, 0.90 0.85 and 0.80). Using each cutoff, we convert the correlation matrix into an adjacency matrix and construct an undirected graph with the greedy clustering algorithm (Clauset *et al.*, 2004). During the clustering process, we reserve the order of neighboring variants and require the largest within-cluster gap <100 consecutive variants. Each cluster in the graph represents an LD block for a specific (r^2) cutoff, containing correlated variants interspersed with less than 100 uncorrelated variants. We then organize these blocks into a hierarchical structure with 8 levels (Fig. 2.1a). At the leaf level (G^0), each node represents a single variant. At higher levels in a sequential order (G^1, ..., G^6), each node represents variants belonging to an LD block with $r^2 > 0.999$, 0.98, 0.95, 0.90 0.85 and 0.80, respectively. The root level (G^7) has a single node containing all variants.

2.2.2 TreeMap Framework

TreeMap takes a 3-layer nested design to remove uninformative variants and reduce redundancies among informative variants progressively (Fig. 2.1b). At the outer layer, tree-guided penalized regression selects groups of variants (internal nodes in G) or individual variants (i.e., leaf nodes in G) associated with transcriptional changes. At the middle layer, stepwise conditional multivariate tests iterate combinations of variants within each selected node to identify a node-specific optimal solution. At the inner layer, variants selected from the previous layers are aggregated and passed through a Bayesian multivariate analysis to derive a global optimal solution. The final solution satisfies both between-locus sparsity by selecting only a few internal nodes, and within-locus sparsity by selecting only a few individual variants in a node. Below we provide detailed descriptions of each layer.

Outer Layer

We formulate the selection of causal variants from a genomic region with LD structure as a sparse learning problem under graph constraints. Specifically, given a feature matrix X with n rows and m columns, a response vector Y of length n, and a hierarchical relationship G of features in X with d levels, we will learn a linear model $Y = X\beta + \epsilon$ that solves

$$\min_{\beta}(\sum(Y - \beta X) + \lambda \sum_{i=0}^{d} \sum_{j=1}^{m_i} w_j^i \|\beta_{G_j^i}\|) \tag{2.1}$$

where β is a vector of individual variants, $\beta_{G_j^i}$ s the vector of coefficients of variants belonging to a node G_j^i, λ is the regularization parameter, and w_j^i is the weight of each node in group G_j^i. We compute w_j^i as

$$w_j^i = \frac{\sqrt{k} + \bar{f}}{\frac{1}{k}\sum_{q \in G_j^i} s_q} \tag{2.2}$$

where l is the number of variants in the group, $\bar{f}$ is the average minor allele frequency, and s_q is a user-specified functional impact score of a variant q in the group (large values for functionally important variants, default value = 1).

The sparsity (i.e., the number of variants with non-zero β values) of the solution to equation 2.1 is controlled by λ, such that a larger λ value leads to fewer selected variants. In practice, choosing the most appropriate value of λ is mostly subjective. To address this problem, we test a range of values with bootstrap samples. The top 5% most frequently selected variants (receiving non-zero β values in bootstrapped samples) are informative. The β value of an internal node is the average of its member variants. The top 5% internal nodes receiving non-zero β values are also informative. We denote the set of variants selected at this layer as S_1.

Middle Layer

For each informative internal node, we perform a stepwise conditional analysis to find a set of variants in S_1 with non-redundant information. Specifically, given a node containing a set of variants V, we first fit a linear regression model for each and every member variant q as

$$Y = \beta_0 + \beta X_{q \in V} + \epsilon \tag{2.3}$$

Among all member variants passing a statistical threshold (i.e. Bonferonni-corrected p value < 0.05 and explained residual $> 1\%$), we choose the variant with the smallest p-value as the primary variant. Next, conditional on this primary signal, we test each remaining variant by fitting a linear regression model on the residual and identify-

ing the variant with the smallest p-value. We repeat this process until exhausting all member variants or no remaining variants passing the statistical threshold. We then aggregate variants selected from this procedure with variants in S_1 that do not belong to any informative internal nodes, and map them into nodes at the G^6 level (i.e., $r^2 > 0.8$). Within each node, we iterate all combinations of one or two variants to fit a linear regression model and compute the Akaike information criterion (AIC) values

$$AIC = 2K - 2\ln L \tag{2.4}$$

where K is the number of variants and L is the likelihood of the fitted model. We select the variants giving rise to the smallest AIC value and denote this set as S_2.

Inner Layer

If S_2 contains no more than 10 candidate variants, we perform an exhaustive search for the best linear model with an arbitrary number of variants based on the Bayes factor (BF). We define M as a multivariate linear model with selected variables and M_0 as a null model with no independent variables. By giving equal prior probabilities to M and M_0, the BF is

$$BF = \frac{Pr(X, Y|M)}{Pr(X, Y|M_0)} = \frac{Pr(M|X, Y)Pr(M_0)}{Pr(M_0|X, Y)Pr(M)} = \frac{Pr(M|X, Y)}{Pr(M_0|X, Y)} \tag{2.5}$$

The set of variants giving rise to the largest BF value constitutes the lead variants of the credible set. If S_2 contains more than 10 candidate variants, we use backward stepwise selection based on AIC values as in equation 2.4 to identify lead variants. Using each lead variant as an anchor, we scan S_1 for tagging variants with $r^2 > 0.5$ linked to the lead variant. We define an eQTL as a lead variant with its tagging variants ranked on r^2 values. The final credible set may contain multiple loci.

Estimating effect sizes

After we derive a final credible set for a gene, we build a linear regression model

$$Y = \beta_0 + \sum_i \beta_i X_i + \epsilon \tag{2.6}$$

where Y is the transcript abundance, X_i is the lead eVar of the ith eQTL and β_s are effect sizes, and ϵ represents error. For each X_i, we test the null hypothesis of $\beta_i = 0$ and use the p-value to represent the statistical significance of the corresponding eQTL. We consider the eQTL with the best p-value as the primary locus and the remaining eQTL as auxiliary loci.

2.2.3 Simulation Data

We used an established approach (Zeng *et al.*, 2017) to simulate gene transcription controlled by one to ten causal variants. Given a randomly picked human gene, we retrieved genotypes X of all variants located in the 200 kb upstream region of its transcription start site from the 1000 Genomes Project phase 3 data (Consortium *et al.*, 2015). From among these variants, we picked h random variants as causal variants, and assigned each causal variant i an effect size $\beta_i = \sqrt{\frac{VE_i}{2\rho_i(1-\rho_i)}}$ where ρ_i is the minor allele frequency and VE_i is the variance explained. We allowed VE_i to take a random value from a uniform distribution $unif(0.02, 1)$. We then simulated gene transcript abundance $Y_i = \sum_{i=1}^{h} \beta_i X_i + \epsilon$ where ϵ is the environmental noise following a normal distribution $norm(0, 1)$. To simulate causal variants in functional genomic regions, we used a pre-compiled DNase I hypersensitive sites (DHS) map (Trynka *et al.*, 2013) that combined DHSs in 217 cell types from the ENCODE and the Roadmap projects. We randomly selected variants inside and outside DHS as causal variants. Variants inside DHSs received $S_q = 1$, whereas variants outside

DHSs received $S_q = 0$.

2.2.4 GTEx Datasets

We downloaded RNA-seq and genotype data of 123 brain hippocampus samples and 274 transverse colon samples from the GTEx data portal (v7, mapped to the hg19 reference genome). Transcript abundance quantified as Transcripts Per Kilobase per Million mapped reads (TPM) were available for 23,725 genes in brain and for 24,423 genes in liver. Following the recommendations from the original GTEx study (Consortium *et al.*, 2017), we adjusted TPM values for covariates using multivariate linear regression. For brain tissues, these covariates include 3 genotyping principal components, 15 PEER factors (Stegle *et al.*, 2012), sequencing platform, and sex. For colon tissues, these covariates include 3 genotyping principal components, 30 PEER factors, sequencing platform, and sex.

To obtain genotype data, we downloaded the vcf file that contained high-quality calls from whole-genome sequencing experiments. These variants include single nucleotide variants and short indels that have passed stringent filters (PASS flag, GQ20, Hardy-Weinberg equilibrium, etc.). For each gene, we applied TreeMap to common variants (minor allele frequency $MAF > 0.05$) located inside the region from 2 Mbps upstream of the transcription start site (TSS) to 2 Mbps downstream of the transcript end site (TES). On average, each gene had 8,199 common variants. We built a hierarchical tree of these variants using the method described above and applied TreeMap to each gene.

2.3 Results

We compared the performance of TreeMap with stepwise conditional analysis, DAP, CaVEMaN, and CAVIARBF using the simulation dataset. We then applied to

GTEx data of brain hippocampus samples and transverse colon samples.

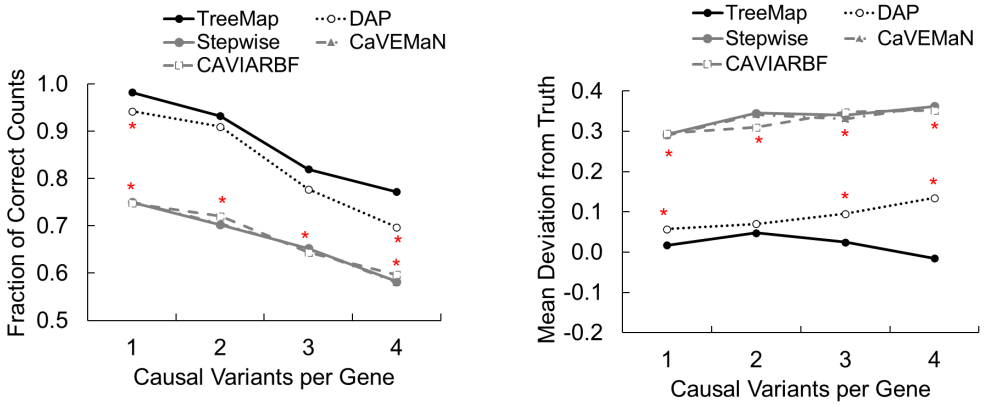
2.3.1 Performance on Simulation Data

Mapping independent causal variants

We randomly sampled 400 genes from the human genome and simulated 1, 2, 3, and 4 causal variants for each gene. We required that r^2 values between all pairs of causal variants of a gene were less than 0.1. These simulations represented genes with only independent cis-acting eQTL.

We first examined if each method reported the correct number of independent eQTL. When a gene had a single causal locus, TreeMap found the correct number 98% of the time, which was significantly higher than DAP (94%, two-proportion test P=0.003), stepwise analysis, CaVEMaN, and CAVIARBF (75%, P=10^{-21}, Fig. 2.2a). Because CaVEMaN and CAVIARBF used statistics from stepwise analysis as priors, it is not surprising that these three methods reported similar results. As the number of independent causal loci per gene increased, the accuracies of all methods decreased linearly. However, the accuracy of TreeMap remained as the highest in all scenarios. For genes with four independent causal loci, TreeMap still made correct predictions 77% of the time, whereas the accuracies of DAP dropped to 70% (P=0.01), and the other three methods dropped to 58% (P=10-9). When these methods made wrong predictions, they tended to over-estimate the number of independent causal loci, with stepwise analysis showing the largest deviations and TreeMap showing the smallest deviations (Fig. 2.2b).

We then examined the sizes of credible sets (i.e., number of putative causal variants at a locus) reported by each method. A credible set contains a lead eVar and additional linked eVars. Small credible sets help narrow target candidate variants



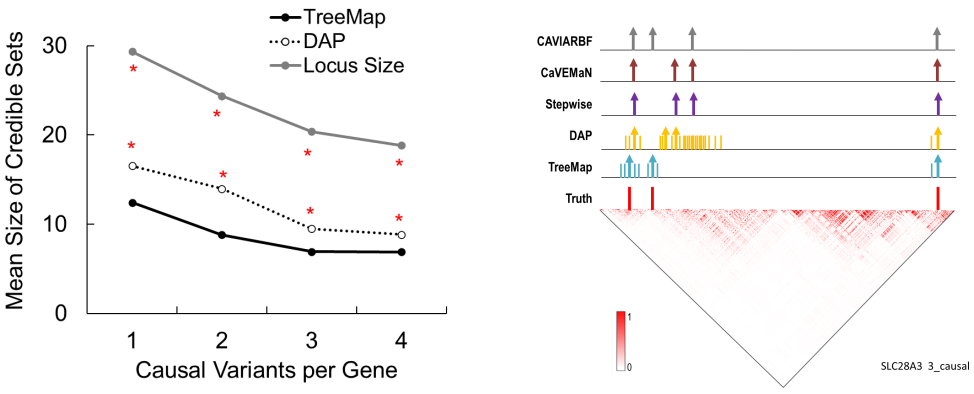
(a) Fraction of genes with correctly predicted numbers of eQTL.

(b) Mean deviation of the predicted and the true number of eQTL.

Figure 2.2: Performance on Mapping Independent Causal Variants by TreeMap, DAP, Stepwise Conditional Analysis, CaVEMaN, and CAVIARBF. * Indicate Significant Differences of Performance Between TreeMap and Corresponding Methods (P$<$ 0.05).

and are thus preferred. On average, a causal variant was linked to 23 variants with $r^2 \geq 0.8$. Among these linked variants, TreeMap selected only 37%-42% to include in credible sets, whereas DAP kept 51%-61%. Therefore, the credible set of TreeMap was significantly smaller than that of DAP (all paired t tests P$<$0.05, Fig. 2.3a). Because credible sets reported by stepwise analysis, CaVEMaN and CAVIARBF contained only lead eVars, we did not include them in the analysis.

A representative example was simulations of 3 causal variants upstream of the SLC28A3 gene (Fig. 2.3b). TreeMap predicted 3 eQTL correctly. At each locus, the lead eVar matched the causal variant. All other methods predicted one extra eQTL. DAP, stepwise analysis, and CaVEMaN found only one causal variant. Although CARVIARBF recovered all three causal variants, it was unable to remove the extra spurious eQTL.



(a) Mean size of credible sets

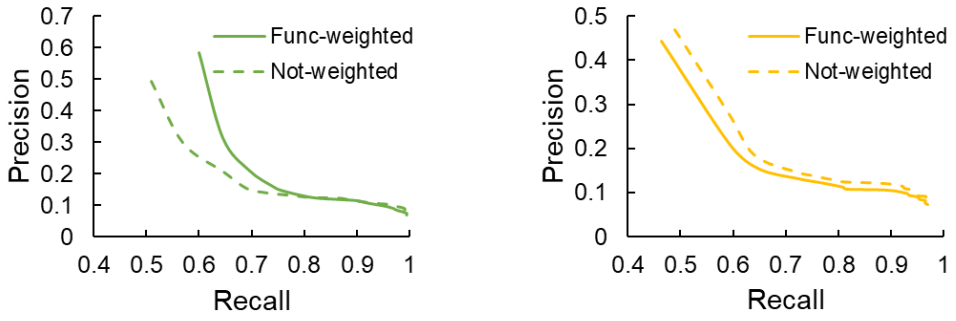
(b) An example with 3 simulated causal variants (SLC28A3 gene)

Figure 2.3: Performance on Mapping Independent Causal Variants by TreeMap, DAP, Stepwise Conditional Analysis, CaVEMaN, and CAVIARBF.

Prioritizing functionally important variants

TreeMap takes user-specified functional impact scores to prioritize variants. To test this functionality, we used DHS annotations as functional evidence. Variants at DHSs received S_q=1, whereas variants outside DHSs received S_q=0. We simulated 200 genes each with a single causal variant at a DHS. As expected, using DHS functional scores led to significantly better performance than using default functional scores (i.e., not-weighted). It showed a better precision-recall curve and identified significantly more causal variants as lead eVars (chi-squared test P=0.02 (Fig. 2.4a)).

To test if inappropriate functional scores decrease the performance, we simulated 200 genes each with a single causal variant outside DHSs. In these data, using DHS functional scores led to a worse precision-recall curve than using equal functional scores, and identified fewer causal variants as the lead variants, although the difference was not statistically significant (P =0.08).



(a) Causal Variants located at DHSs (b) Causal Variants outside DHSs

Figure 2.4: Precision-Recall Curves of TreeMap Using DHS-derived Functional Scores.

Computational efficiency

We simulated 2,000 genes, each with 1 to 10 causal variants. We distributed these causal variants randomly in the 200 kbps upstream regions of a gene. Each gene had an average of 1,712 variants genotyped in 1,835 samples. The pairwise linkages of these causal variants covered the full range of r2 values from 0 to 1. We then executed each method as a single-threaded process on a Dell laptop computer with an Intel® Core™ i7-7600 CPU at 2.80 GHz and 16GB RAM.

Stepwise analysis was the most efficient method, taking an average of 9.0 seconds (secs) to analyze a gene with a single causal variant, and 38.3 secs to analyze a gene with 10 causal variants (Fig. 2.5). The computational efficiency of TreeMap and DAP were moderate. To analyze a gene with only one or two causal variants, DAP took a shorter time than TreeMap (mean CPU time = 24.8 – 37.3 secs for DAP, and 104.1 – 116.3 secs for TreeMap). However, when the number of causal variants increased, the CPU time of DAP increased exponentially. For a gene with 6, 8 or 10 causal variants, DAP took an average of 965.2, 1610.6 and 5034.0 secs (16.1 minutes to 83.9 minutes) to analyze it. The CPU time of TreeMap was stable, increasing only to 162.6, 178.0 and 189.0 secs (2.7 to 3.2 minutes) in these cases. CaVEMaN and CAVIARBF were

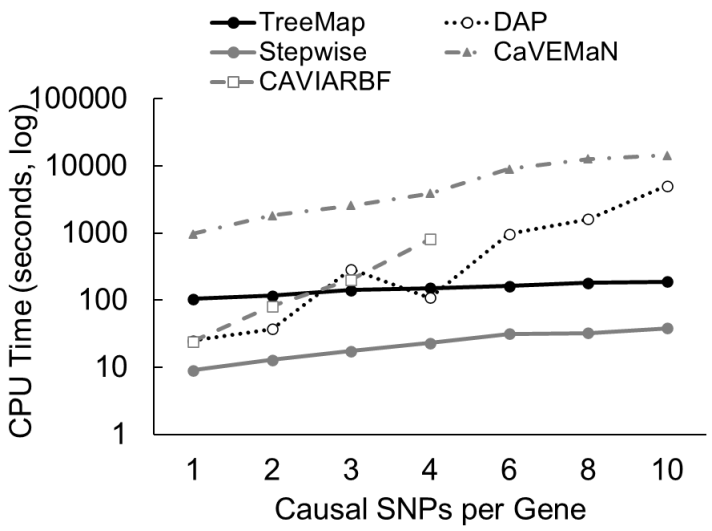
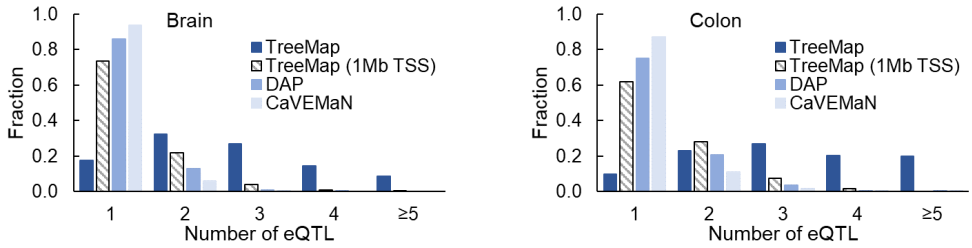


Figure 2.5: Average CPU Time (Seconds in Log Scale) to Analyze One Simulated Case.

the slowest. Because CAVIARBF performs exhaustive search, it was computational prohibitive to apply this method to genes with six or more causal variants.

2.3.2 Applications to GTEx data

We retrieved genotype and transcriptome profiles of 123 brain hippocampus samples and 274 transverse colon samples from the GTEx data portal. There were 23,410 genes expressed in at least 10% of the brain samples and 17,065 genes expressed in at least 10% of the colon samples. For each gene, we retrieved genetic variants in a large genomic region that spanned from 2 Mbps upstream of the transcription start site (TSS) to 2 Mbps downstream of the transcription end site (TES). After removing rare variants with MAF$<$ 5%, each gene had on average 8,281 genetic variants in this region. For each gene, we applied TreeMap to organize variants into a hierarchical tree based on pairwise r^2 values, and to identify eQTL and putative causal variants guided by the tree. To correct for multiple comparisons, we required that the primary



(a) Brain (b) Colon

Figure 2.6: Fractions of Genes with Single or Multiple eQTL Identified by TreeMap, TreeMap Restricted to 1Mb Flanking TSS, DAP, and CaVEMaN.

eQTL locus of a gene had a p-value $< 10^{-6}$ corresponding to a false discovery rate of approximately 0.01 (i.e., 10^{-6} x 8,281). For auxiliary loci, we applied a lenient cutoff of p-value < 0.01 because these were post hoc tests after a significant primary eQTL was identified (Kim, 2015).

We detected eQTL of 4,950 genes in brain samples and eQTL of 4,636 genes in colon samples. In both tissues, the majority (69%-73%) of genes had two to four eQTL. (Fig. 5A-B). This was in contrast to DAP and CaVEMaN results that reported single eQTL for most genes. To examine if this difference was due to TreeMap scanning longer regions than the other two methods (>4 Mb vs. 1 Mb), we restricted TreeMap to 1 Mb flanking TSS. Indeed, we found single eQTL for most genes. Therefore, the additional eQTL identified by TreeMap were owing to its capability to analyze variants beyond 1 Mb to TSS.

2.4 Discussion

With increasing sample sizes for eQTL mapping it has become apparent that the majority of all genes have a complex pattern of regulation influenced in cis by multiple SNPs. Fine mapping of the causal variants is constrained by the high degree of LD covering most regulatory regions, and high levels of polymorphism such that

credible intervals average 100 sites or more (Zeng *et al.*, 2019). Three broad approaches to dealing with this complexity are being developed: stepwise conditional regression, Bayesian dimensionality reduction, and haplotype-based modeling. The method introduced in this study, TreeMap, combines elements of the latter two.

An important aspect of haplotype-based methods is the heuristic definition of haplotypes. Perhaps the most rigorous procedure uses the four-gamete test to identify minimal length haplotypes by virtue of inferred recombination events. A genome-wide association mapping method based on this approach, HaploSNP (Sargent *et al.*, 2016), explains much more of the variance per locus. However, because haplotypes are greatly susceptible to biases introduced for example by population structure and do not have base pair resolutions, HaploSNP has not been widely adopted for fine mapping. Instead, we here propose TreeMap, a hierarchical approach based on successive LD thresholds. Causal variants are assumed to be embedded in LD blocks although the extents of linkage are unknown. Transcriptional effects are gleaned from comparison of the likelihoods of models for blocks defined by varying LD thresholds. This algorithm is thus independent of cladistic methods for assembly of cladograms with ad hoc thresholds that may have hampered adoption of earlier iterations of haplotype-based approaches (Sargent *et al.*, 2016; Templeton *et al.*, 2005).

Using extensive simulation, we show that TreeMap modestly, yet significantly, outperforms representative alternative multisite eQTL mapping algorithms (stepwise conditional regression, and DAP) in several key regards. First, it recovers more independent variants, particularly as the complexity of multisite regulation increases. Second, it reduces the size of the credible interval as assessed by improvement in the precision-recall curve. Third, it recovers more causal variants under LD. Furthermore, since the method is computationally far less demanding than even the fastest Bayesian approach, DAP, it is possible to scan 4Mb around the transcription start site, and this

doubling of the potential regulatory region led to the discovery of multiple hitherto unrecognized distal eQTL in the GTEx dataset.

2.5 Limitations and Future Direction

There remain several limitations to be addressed. Like the other methods, performance drops as the number of independent causal variants in an eQTL increases, particularly if they fall within intervals of high LD. Under soft selection scenarios, it may be expected that regulatory regions will harbor more than one variant influencing gene expression, with multiple signals embedded in a haplotype. Variants that have opposing signs of effect will tend to reduce the overall signal. Methods for multisite mapping of tightly linked causal variants need to be further explored. One strategy is to incorporate functional evidence from ENCODE or evolutionary conservation, or computationally predicted impact scores into the mapping algorithm (Cannon and Mohlke, 2018; Yang *et al.*, 2017). TreeMap has a built-in mechanism to prioritize variants on functional scores, though the performance depends on the appropriateness of functional scores. For this reason, we leave the choice of these scores to users. It may be necessary to combine functional evidences from multiple domains to derive a composite score (Guan *et al.*, 2020; Liu *et al.*, 2019). Finally, because the definition of credible sets in TreeMap is not associated with a probability, its statistical interpretation is not as straightforward as in Bayesian-based approaches.

Our simulations are restricted to an effect size with beta extracted from unif(0.02, 1) and error range of norm(0, 1). Further downstream analysis of varying signal-to-noise ratio needs to be explored. Hence, the users of our algorithm need to be cautious about expanding the simulations and using our algorithm for other scenarios.

TreeMap, and future improvements of incorporating prior biological knowledge in fine mapping algorithms, facilitate discoveries of regulatory variants from genome-

association studies and whole-genome sequencing studies as well. The capability of searching long genomic regions makes it a promising approach to identifying novel distal regulatory variants underlying human diseases and other health-related genotypes.

Chapter 3

DEEPCORE: INTERPRETABLE MULTI-VIEW DEEP LEARNING APPROACH FOR DETECTING CO-OPERATIVE REGULATORY ELEMENTS

3.1 Background

3.1.1 Gene Expression is a Molecular Phenotype

Pathogenic noncoding variants exert their functional impact via disrupting transcription regulatory elements that subsequently alter gene expression. Therefore, a promising approach to fine mapping causal noncoding variants is via modeling the relationship between noncoding variants and gene expression. An initial step to this challenging task is the accurate prediction of tissue-specific gene expression. Once this model is in place, it can be used to identify regulatory elements and predict transcriptional impact of genetic variants that vary across tissue and cell types.

Gene expression prediction has been studied for more than a decade. Early techniques performed a reverse analysis where they first identified genes based on the expression level and attempted to find signature defining regulatory elements (Brāzma *et al.*, 1998; Liu *et al.*, 2000; Roth *et al.*, 1998; Sinha and Tompa, 2000). Current techniques for gene expression prediction can be categorized into two groups:

- Using genetic sequences
- Using epigenetic signals like TF occupancy (TFO) and Histone Modifications (HMs).

Using Genetic Sequences

Segal *et al.* (2008) developed a thermodynamic model which take the regulatory sequences and position-specific scoring matrix (PSSM) of the factors as input to compute occupancy distribution (probability distribution of the number of factors on a given sequence) which was then converted to expression level using a logistic function. The final expression level is the weighted sum of each contribution. Segal *et al.* (2002) built a unified probabilistic relational model to predict gene expression levels on the yeast genome. They defined 1000bps upstream of the yeast genome as the promoter regions and used the sequences in these regions for prediction. Their unified model comprises of two parts: predicting whether transcription factor regulates the gene given the promoter sequence, and predicting expression level of a gene given the regulatory events. Both of these parts are trained in a single model to achieve maximum efficiency.

Another study on yeast involved using a Bayesian network to infer gene expression profiles from the sequence elements (Beer and Tavazoie, 2004). They first used k-means clustering of microarray expression data to identify 49 clusters of co-expressed genes for given environmental stress and cell cycle using 800bps upstream sequence of the genes. They then build a sequence-to-gene bayesian network to predict the probability of a gene belonging to a particular expression pattern. Yuan *et al.* (2007) used the same clustered data from above. They used motif matching scores of 666 transcription factors to build Naïve Bayes classifier for each bin to predict the cluster label representing different gene expression patterns.

The most recent approach is ExPecto (Zhou *et al.*, 2018), a deep learning based model, that used DNA sequences of 40 kbps around the TSS of the genes to predict 2002 different epigenetic markers which were then given to a regularized linear model

to predict the expression patterns of the genes in various tissues.

Using Epigenetic Signals

As we have seen earlier, epigenetic markers like TFO and HMs play a major role in gene regulation. Various studies have attempted to predict the gene expression using epigenetic summary tracks. A vast majority of the techniques used features related to summary statistics derived from ENCODE epigenetic tracks but not low-level epigenetic information to predict gene expressions. For example, Bussemaker *et al.* (2001) used the count of the motifs recognized by TFs as the predictor, Keleş *et al.* (2002) used the weighted sum of the motif occurrences based on distance from the gene and built a stepwise linear regression model, Conlon *et al.* (2003) used motif scores as predictors to build a stepwise regression that detected significant groups of motifs affecting gene expressions.

Ouyang *et al.* (2009) used TFO measured by Chip-seq technology to predict gene expressions. They constructed TF association strength (TFAS) from each TF for each gene by computing the weighted sum of chip-seq signals where the weights corresponded to the distance of the signal to the gene. Principal component analysis (PCA) was applied to account for patterns in TFAS and also to reduce the dimension which was then used as covariates in the linear regression model to predict variations in gene expression. Costa *et al.* (2011) studied the effect of HMs and TFO on cell differentiation and cell-specific gene expression. They looked at HMs and TFO for four T-cell types and built a linear mixture model with two regression models to predict the cell-specific gene expressions. They looked at the regression coefficients for various histone modifications for the two models. In both the models, they found that H3K4me3 had the largest influence among the highly expressed genes and H3K27me3 consistently showing up among genes with low expression.

Cheng *et al.* (2011) divided regions of 4k base pairs(bps) around the transcription start site(TSS) and transcription termination site(TTS) for multiple transcripts of various genes in C. elegans worms into bins of 100bp each and computed average signals value for 13 chromatin features involving histone modifications and various binding factors. They used SVM to each bin and predicted the expression level of each gene. They also used clustering and correlation analysis to study the relationship between gene expression and chromatin features and found two clusters of genes that are characterized by two different sets of chromatin features. Cheng and Gerstein (2012) modeled the relationship of TFO and HMs with gene expression among mouse stem cells. They used a similar binning strategy as their other work. They applied a two-layered Support Vector Regression (SVR) to predict gene expression levels. Dong *et al.* (2012) proposed a two-step quantitative model for gene expression prediction from chromatin features. They used a similar strategy as (Chen *et al.*, 2016; Cheng and Gerstein, 2012) in binning the data. However, they used a novel strategy of 'best bin' (bin having the maximum absolute correlation with gene expression) and used this to build a two-step predictive model. In the first step, they built a Random Forest classification model to classify genes into 'on' or 'off' state. They then built a linear regression model to quantify the gene expressions for the genes classified as 'on'. Schmidt *et al.* (2017) built a linear regression model using the computed transcription affinities for various TFs across open chromatin regions to predict gene expression.

DeepChrome (Singh *et al.*, 2016) was the first deep learning method used for gene expression prediction from HMs. They applied a convolutional neural network model on binned 5kbps regions around the TSS of a gene to predict the gene expression level using five core HMs. As the convolution learns the combinatorial interactions among different histone marks, their model achieved higher accuracies compared to all the other models. They also built a visualization technique to display the interactions

between HMs.

Drawbacks of the existing approaches

Although all the above-mentioned techniques have shown an association between epigenetic information, sequences, and gene expressions, they have the following limitations:

- Most of the studies are based on organisms other than humans and only a handful of them has attempted to find the impact of genetics and epigenetics on gene regulation.

- The vast majority of these studies have moderate accuracies which can be due to either insufficient data or the use of linear models

- Many studies use bins and build separate models for separate bins and fail to look at the relationship and combinatorial interactions among these bins.

- Although DeepChrome [37] addressed the issues of combinatorial interactions among multiple bins, the strategy of binning and averaging the feature intensity for each bin suffers from boundary effect and weakens the predictive power of the features used mainly because the peaks can be assigned to multiple adjacent bins. We have seen both genetic sequences and epigenetic marks play a vital role in gene expression. But, to the best of our knowledge, none of the studies have attempted to combine sequences and epigenetic information in predicting gene expressions.

- All the existing methods along with gene expression prediction also showed us which of the epigenetic information was more useful in predicting or was related to the gene expression and none of them focused on the actual regions around

the TSS to find function-related patterns. Even the methods that used the binning strategy failed to take a close look at each bin and find relationships among them.

- No method considers cooperation of regulatory elements that is fundamental to transcriptional control

3.1.2 Extracting Useful Features from Deep Learning Models

Although neural networks have existed for quite some time, the majority has treated and left it as a black box unit. One of the early papers to visualize the important features from the neural network model is a study by Zeiler (Zeiler and Fergus, 2014). They introduced a novel technique called Deconvolutional Network(deconvnet) which reverses the process of convolution. A typical CNN has three major steps: convolution, pooling, and rectification in the same order. The authors took the output of a CNN layer and recreated the input matrix. They first passed it through the rectification and then the unpooling step uses the switches to place the max values at the switch locations. In the deconvolution step, they transpose the filter used in convolution to the pooled maps.

Simonyan *et al.* (2013) used a numerical optimization approach to visualize the convolution models. They generated images representing each class using optimization. They initialized a zero image and used backpropagation to add the mean of the train set images to the zero image. The final image represented a single image representing a given class and has a high class-score as computed by the classification layer. They also looked to find useful patterns in individual images by finding saliency maps of images. They used sensitivity analysis to obtain saliency maps of images computed by the DNN. They looked at the effect of small changes in each

pixel on the network output.

Zintgraf *et al.* (2017) took a different approach in visualizing the important features. They focused on identifying and visualizing the importance of a neuron in a layer or the importance of a layer itself that contributes towards the prediction of an image to a class. They used a novel measure called the weight of evidence(WE) to measure the importance of a unit in prediction decision. WE is the difference in the odds of an image belonging to a class given a particular unit and the odds of an image belonging to a class when the unit is not present. Mahendran and Vedaldi (2015) proposed an inversion technique to visualize the features. They used the encoding at the end of each layer to create an inversion of the image and compare it with the original image. The intuition behind this inversion technique is that the pixels that can be obtained and reconstructed at a given layer are the content of the layer.

All these methods can be categorized into two groups: Backpropagation and Reconstruction. Although there are only a handful of other projects that focus on visualization, they all belong to one of these two categories. All of these methods are applied to images and visual computing. Since the DNN feature visualization is a new and ongoing research topic, its application in the field of bioinformatics is also very limited. DeepChorme (Singh *et al.*, 2016) used numerical optimization to find an image that best represented a particular class and then visualized the image by constructing a heat map of the reconstructed image which was similar to Simonyan *et al.* (2013). DeepMotif (Lanchantin *et al.*, 2016), which classified TFBS based on gene sequences, also applied the same technique as Simonyan *et al.* (2013) to reconstruct the image for each class.

There are three major drawbacks in the current methods for detecting useful potential regulatory elements. Firstly, in visualizing features for gene expression prediction, the fundamental units are a group of sequences rather than each individual

pixels (sequence in our case). As the goal lies in identifying a small region which contributes towards regulation, the existing methods might not work for our data. Secondly, most of the methods focus on class based feature extraction. However, we are interested in each individual gene rather than each class. Also, we use regression for predicting gene expressions which rules out class based feature extraction. Finally, to the best of our knowledge, vast majority of the research is focused on visualizing CNNs and feed forward neural networks. It is harder to visualize RNNs and LSTMs.

To address the limitations, we have developed a novel attention-based deep learning method that identifies regulatory elements by modeling the relationship between genetic, epigenetic, and transcriptional signals.

3.2 The Novel DeepCORE Framework

3.2.1 Data Extraction and Preprocessing

For each gene, we extracted the DNA sequence and histone modification marks for 10Kbps DNA surrounding the transcription start site (TSS, 5Kbps upstream, and 5Kbps downstream).To derive genetic features, we obtained TF-DNA binding preferences for each gene for 50 different TFs from JASPAR (Sandelin *et al.*, 2004). For each TF, we scanned the 10Kbps region around the TFs and encoded TF activity to the genomic region using a binary scheme (1 – presence of TF activity, 0 – no TF activity). If a TF has overlapping binding sites, we sum up the TFBS scores. From this, our genetic features are represented as a matrix where rows represent different TFs and columns represent the genomic regions.

To derive epigenetic features, we used data from ENCODE (Consortium *et al.*, 2004) and RoadMap Epigenomics project (REMC) (Bernstein *et al.*, 2010). To avoid differences in sequence depths, REMC normalized the read counts of all the histone

marks to a depth of 30 million reads. We obtained these normalized read counts from ChIP-seq experiments that measure five histone modification marks that are known to impact gene expression either positively or negatively (Kundaje *et al.*, 2015) as depicted in **Figure 1.3**. For tissue-specific transcript abundance of each gene, we used Reads Per Kilobase of transcript per Million mapped reads (RPKM) from RNA-seq experiments in the REMC project.

We extracted the tissue-specific histone marks and the gen- specific TF binding features for 25 different cell lines. For each cell line, we removed all the genes which did not have signals across all five histone marks or had no TF activity across all the 50 TFs.

3.2.2 DeepCORE Model for Tissue-specific Gene Expression Prediction

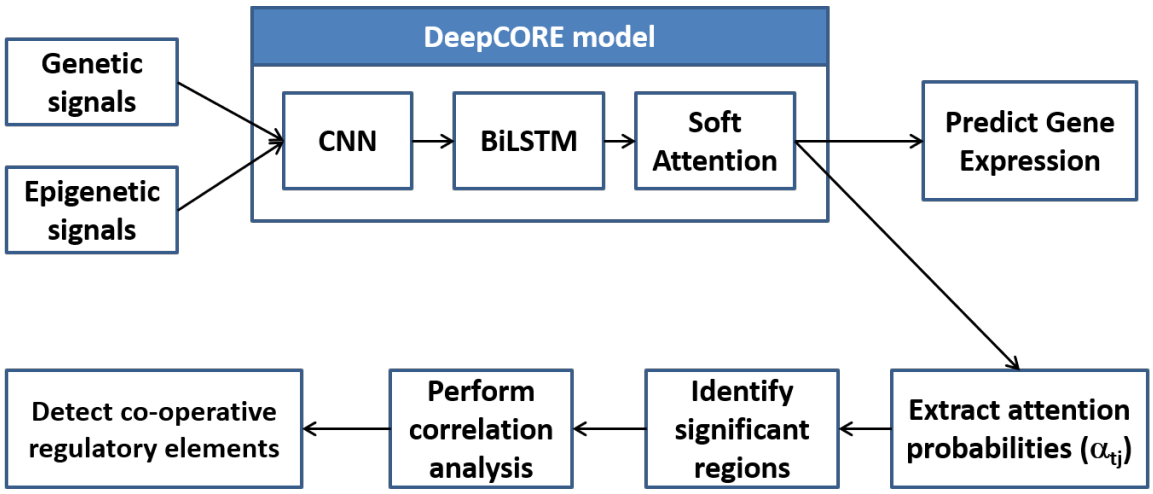


Figure 3.1: DeepCORE Framework

We present a novel multi-view attention-based deep learning model as shown in **Figure 3.1** which takes DNA sequences and the epigenetic marks in the form of HMs as inputs and predicts tissue-specific gene expression levels. The deep learning model contains two separate paths representing two different views that undergo

convolution and lstm operations for sequences and histone modifications separately which differ from the traditional approaches of using combined convolution for a given data (**Figure 3.2**). The main reason for the multi-view is to understand and identify which areas of the genetic sequence and epigenetic sequence are responsible for the gene regulation and also to visualize the important features.

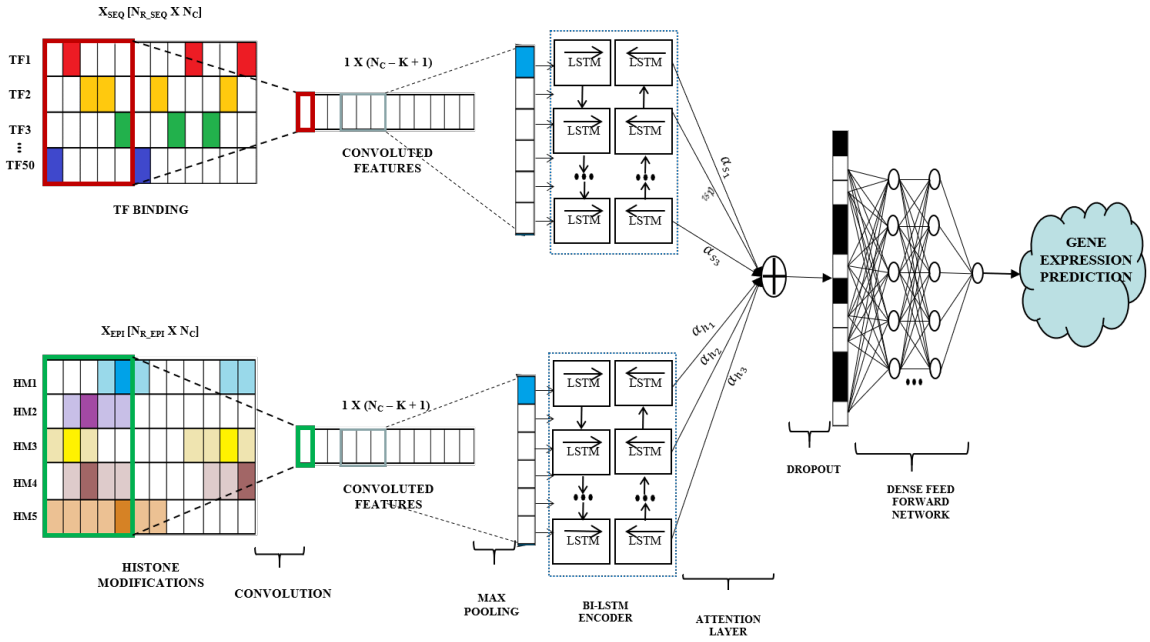


Figure 3.2: DeepCORE Architecture

Convolution layer

In this layer, we use separate convolutions for gene sequences and epigenetics. It consists of N_{cnn} filters with size k and stride 1 for both sequence convolution and epigenetic convolution. This performs sliding window across all 10Kbps and learns important features from them.

$$Z_{seq_m,i} = \beta_i + \sum_{j=1}^{K} W_{i,j} X_{seq} \tag{3.1}$$

Equation 3.1 represents convolution for m_{th} sliding window for i_{th} filter, where $I \in \{1, \ldots, N_{cnn}\}$ and $m \in \{1, \ldots, N_C - K + 1\}$. Similarly, we also compute convolution features for epigenetic marks which is represented by Z_{epi}. The output of the convolution layer is passed to ReLU function which forces the negative weights to zero. ReLU is the most widely used activation function which accounts for non-linearity and interactions in the model. ReLU is defined using equation 3.2.

$$f(x) = max(0, x) \tag{3.2}$$

In order to eliminate all the duplicate and irrelevant features, pooling is generally used. We used maxpooling which selects the max value for a given window size. ReLU output which is a $N_{cnn} * (N_c – K + 1)$ matrix is given to max pooling over non-overlapping interval of length p. We will end up with $2 * (N_c – K + 1)/p$ vector of outputs for filter f_i, $i \in \{i = 1, ..., N_{cnn}\}$ where one row is for sequence convolution and another for epigenetic convolution.

BiLSTM layer

Although convultion catches the various patterns within the given genme regions, it still fails to consider the interaction between different regions. As we know that enhancers and promoters which are Kilobase pairs away from each other interact to regulate the gene transcription, we wanted to catch these interactions. RNNs are widely used to capture such long term dependencies and are exensively used in natural language processing tasks like part-of-speech tagging, named entity recognition (NER), and other tasks. They contain units to store signals from previous outputs for making current decisions. However, it suffers from vanishing gradient problem which makes it incapable of capturing long-term dependencies (Bengio *et al.*, 1994). To address this problem, GRU (Cho *et al.*, 2014a) and LSTM (Hochreiter and Schmidhuber,

1997) architectures were introduced and found great success in capturing sequence-based tasks with long term dependencies. AN LSTM cell contains three gates: input gate(i), forget gate(g), and output gate(o) controlling the amount of information to pass between input and output which includes hidden states. These states can be calculated using the following equations:

$$f_t = \sigma(W^f.x_t + U^f.h_{t-1} + b^f) \tag{3.3}$$

$$i_t = \sigma(W^i.x_t + U^i.h_{t-1} + b^i) \tag{3.4}$$

$$o_t = \sigma(W^o.x_t + U^o.h_{t-1} + b^o) \tag{3.5}$$

where W and U are weight matrices and b represents bias which are randomly initialized during the training phase. σ is the sigmoid activation function. Other functions like tanh and ReLU can also be used depending upon the task at hand. A candidate output g is computed using equation 3.6

$$g = \tanh(W^g.x_t + U^g.h_{t-1} + b^g) \tag{3.6}$$

At each time interval t, the cell output state, c_t, and the output, h_t, is computed using the following equations.

$$c_t = f_t \odot c_{t-1} + i_t \odot g \tag{3.7}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.8}$$

where $\odot$ indicates element-wise multiplication. The final output of the LSTM layer is a vector of all the outputs: $Y_t = [h_1, ..., h_t]$. Bidirectional LSTM (BiLSTM) have proved to work better than the traditional LSTMs (Graves and Schmidhuber, 2005; Graves *et al.*, 2013; Fan *et al.*, 2014). BiLSTM captures long range dependencies from

both input directions. It contains two separate LSTM units, forward and backward units, and the two hidden layers are connected to the same output layer as shown in

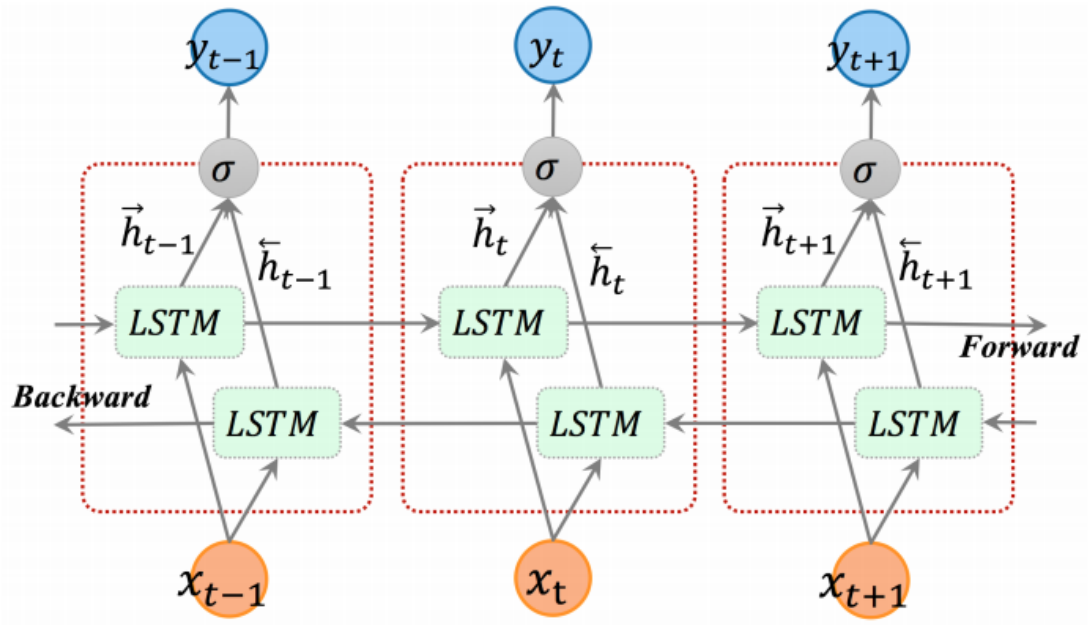


Figure 3.3: Bi-directional LSTM Architecture (Cui *et al.*, 2018)

Attention layer

A major drawback in LSTMs is that as the input sequences gets longer, the more difficult it becomes for the hidden states to capture the context and hence the performance drops (Cho *et al.*, 2014a; Koehn and Knowles, 2017). As more and more updates happen to the same vector, it tends to forget the earlier input updates. For the machine translation task, this was problematic and hence encoder-decoder approach was introduced (Cho *et al.*, 2014b) which is shown in Fig. 3.4.

In this approach, the encoder is a traditional BiRNN model and decoder predicts the next word based on the context vector c and all the previously predicted words as per equation 3.9

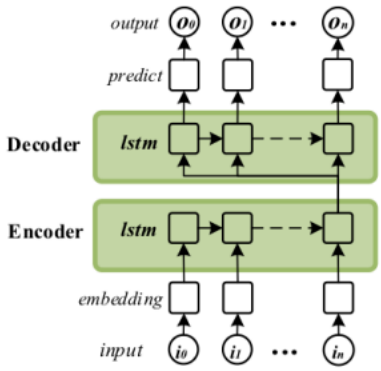


Figure 3.4: RNN Encoder-Decoder Architecture (Cho *et al.*, 2014b)

$$p(y) = \prod_{t=1}^{T} p(y_t|y+1, ..., y_{t-1}, c) \tag{3.9}$$

where $y = (y_1, ..., y_T)$. The conditional probability is given by

$$p(y_t|y+1, ..., y_{t-1}, c) = g(y_{t-1}, s_t, c) \tag{3.10}$$

However this model also suffered the same drawbacks of LSTMs. Bahdanau *et al.* (2014) introduced attention mechanism to address this problem. The idea behind this mechanism is instead of using only the last state for decoding, they used all the hidden states in making the decisions. The difference of this approach with the encoder-decoder architecture is in the decoder portion. The decoder is computed using the following equations.

$$p(y_t|y+1, ..., y_{t-1}, X) = g(y_{t-1}, s_t, c_t) \tag{3.11}$$

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{3.12}$$

Here s_t is the RNN hidden state at timestamp t. c_t, context vector, is a weighted sum of all the annotations (all hidden states from the encoder, $h_1, ..., h_T$).

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j \tag{3.13}$$

$$\alpha_{tj} = \frac{exp e_{tj}}{\sum_{k=1}^{T} exp(e_{tk})} \tag{3.14}$$

$$e_{tk} = a(s_{i-1}, h_j) \tag{3.15}$$

where e_{tk} is an alignment model which is mostly a feed-forward network, α_t is the weight of each annotation which is a softmax function. The attention model is summarized in Fig. 3.5

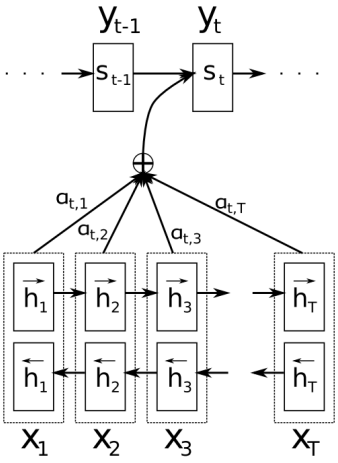


Figure 3.5: Attention Mechanism (Bahdanau *et al.*, 2014)

We modified the regular attention model in two ways. Firstly, the softmax function suffers from non-sparsity in the probability distribution. As we want to pay attention to very important regions contributing towards the gene regulation, we are interested in having a sparse probability distribution. We used a sparsemax function (Martins and Astudillo, 2016) in place of softmax for the sparsity of probability distribution which is given by

$$sparsemax(z) = \underset{p \in \Delta^{K-1}}{\arg\min} \|p - z\|^2 \tag{3.16}$$

where Δ^{K-1} is a $K-1$ dimensional simplex.

Dropout layer

Dropout is applied to the max pooled vector which randomly assigns the inputs to zeros with a probability of 0.5. Dropout is usually applied to avoid overfitting as it is an alternative way of using ensemble learning (Srivastava *et al.*, 2014).

Learning Mechanism

The learnt attention is then given to a fully connected network which is then used to predict the continuous gene expression levels. Mean Squared Error loss is defined to detect the error rate which is then fed back to the network through back propagation. We use a variation of gradient descent approach, called Adam (Adaptive Moment Estimation) optimizer (Kingma and Ba, 2014) to train our model. The model is run for 100 epoch. However, early stopping criteria is employed to avoid overfiting.

Rationale for using the multi-view model

ExPecto is built upon a framework where they first use gene sequences around 40 kbps to predict the epigenetic factors and then transform these predictions as features to predict tissue-specific gene expressions (**Figure 3.6a**). This linear chain of inference implicates that epigenetic changes are only possible as a consequence of genetic changes, which is not true biologically. To reflect the true biological information flow, we choose to use a different framework ((**Figure 3.6b**) where epigenetic changes can lead to gene expression changes without genetic changes. We implicate this framework as a multi-view model.

A potential concern of the multi-view model is the increase of complexity as compared to the linear framework in Expecto. The computational complexity of any machine learning model plays a huge role in its performance. Occam's razor suggests that one should choose a simpler model when given a set of models with similar

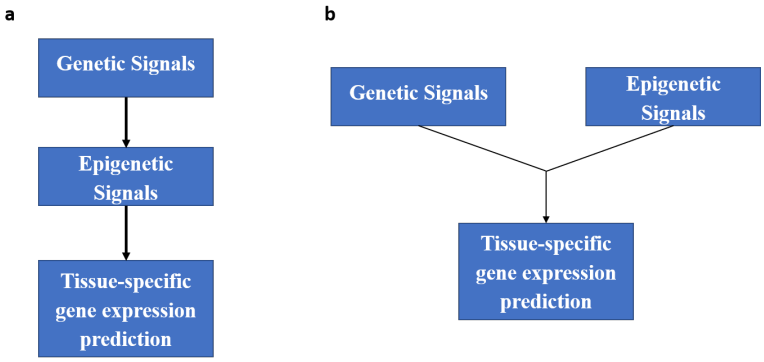


Figure 3.6: Comparison of Modeling Approach Between Expecto(a) and DeepCore(b)

performances. However, our show in the Result section that our model has significantly better performance than Expecto. Furthermore, a simpler model incurs high bias where the model is not able to capture all the informative patterns, such as the over-simplification of the genetic-epigenetic-transcriptional relationship into a linear chain, which risk under-fitting. On the other hand, as the complexity of the model grows, it eventually suffers from high variance where it learns the training sample very well but cannot be generalized and perform poorly on the test samples which often is termed as overfitting. We show in the Results section that our model has good generalizability when applying a model trained on one cell line to data from different cell lines. Furthermore, there needs to be some amount of variation in the features for any model to learn and accurately predict a given task. As There is only a 0.1% difference between any two individuals in terms of genetic sequences meaning the genetic features remain the same and the variations in gene sequences occur only in the form of genetic mutations. As the gene sequence is consistent across all human populations, using genetic signals alone does not allow us to identify the regulatory elements. Therefore, we decided to go with the multi-view model instead of the traditional approach.

3.2.3 *Detect Co-operative Regulatory Elements (CORES)*

There are three major steps for the detection of CORES:

1. Detecting significant genomic regions
2. Correlation analysis
3. Detecting CORES

Each sample undergoes separate convolution and lstm operations using the optimized weights from the DeepCORE model. These operations are hypothesized to learn the adjacent and long term dependencies between various genomic regions. As the CNN operation involves max pooling, these genomic regions are now represented as bins. Once these bins go through the attention mechanism, the model learns to identify the regions that are significant contributors to regulate or repress gene expression levels. We extract the tissue-specific attention probabilities for the bins for each gene and compute the cumulative distribution (CDF) of this attention probability distribution. We then use a normal distribution on these probabilities to extract the p-values and use a cutoff of 0.05. We propose these significant bins (genomic regions) as key functional regulatory elements.

After extracting the highly significant bins for each gene across all the tissue, we end up with a matrix with rows representing different tissues and columns representing the bin probabilities. Pearson's pairwise correlation (Benesty *et al.*, 2009) was then applied to this matrix which gives us the correlation between different regions in the gene. Pearson Correlation is given by equation

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.17}$$

where,

$$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{3.18}$$

The correlation coefficient value ranges between [-1, +1] with

$$\rho_{X,Y} = \begin{cases} > 0, & \text{if one variable increases,the other variable also increases} \\ 0, & \text{no relationship between two variables} \\ < 0, & \text{if one variable increses,other variable decreases} \end{cases} \tag{3.19}$$

We use a combination of correlation and p-value threshold to identify highly correlated bins representing genomic regions that are co-operatively affecting the gene expressions. We first identify all pairs of highly correlated bins and then check if the regions corresponding to these paired bins are at least 1000 bps apart. We propose the bin closest to the TSS to be promoter elements and the region corresponding to it is said to be promoter region. The other distal paired bin is identified as enhancer bin and the region is called the enhancer region. We then create a list of enhancer-promoter region pairs.

3.2.4 Training Procedure

The data is split into training, validation, and test sets. The test set is hidden/held out and is not used until all the optimal parameters are incurred to avoid information leak. Two levels of optimization are performed. The first one is the hyperparameters in the deep learning models. Since the performance of any deep learning model relies highly on the optimization of hyperparameters and the seed values for initialization (Reimers and Gurevych, 2017), we used a grid search approach. Each set of parameters are trained on the training set and then its performance is tested on the validation set. The model with the best performance is used as the optimal configuration and the parameters used in these models are the optimal set. The second

level of optimization is done on the attention mechanism. We tested traditional softmax, softmax with temperature (Hinton *et al.*, 2015), gumbel softmax (Jang *et al.*, 2016), and sparsemax (Martins and Astudillo, 2016) as the activation function to attain sparsity in the attended regions and used the validation set to pick the best activation function. Until this step the test set is never used. Once we obtained the optimal hyperparameters and the attention activation function, we used this optimal model to make predictions on the new held-out test set to account for generalizability of our model.

3.3 Results and Discussion

3.3.1 Multi-view Architecture Improves the Accuracy of Gene Expression Prediction

We split the data into training, validation, and test sets using 80:10:10 ratio. We set aside a set of genes across various tissues for testing. We train our neural network model on various parameters using a grid search and the validation performance to come up with the optimal set of parameters. **Table 3.1** gives all the different combinations of the parameters we tested.

The highlighted set of parameters gave the best validation accuracy and used for all our analysis ($K = 50, N_{CNN} = 50, p = 50, N_{LSTM} = 15, N_{ATTN} = 25, N_{FNN} = 100$). We then use the model with this set of parameters to predict the held our gene set. We used two baseline models to compare the performance of the proposed model: using TF activities alone and using epigenetic alone. Root mean squared error (RMSE) was used as the performance measure for comparison.

Multi-view DeepCORE performs the best when compared to single-view genetic features and single-view epigenetic features in both cell lines producing significantly

Parameter	Values
Convolution filter size (K)	5, 10, **50**, 100, 250
Number of convolution filters (N_{CNN})	**50**, 100
Convolution pool size (p)	5, 10, **50**, 100, 200
Number of LSTM units (N_LSTM)	10, **15**, 25, 50
Number of attention units (N_{ATTN})	10 , **25**, 50
Number of fully connected network neurons (N_{FNN})	50, **100**, 200, 500

Table 3.1: Different Hyperparameter Settings

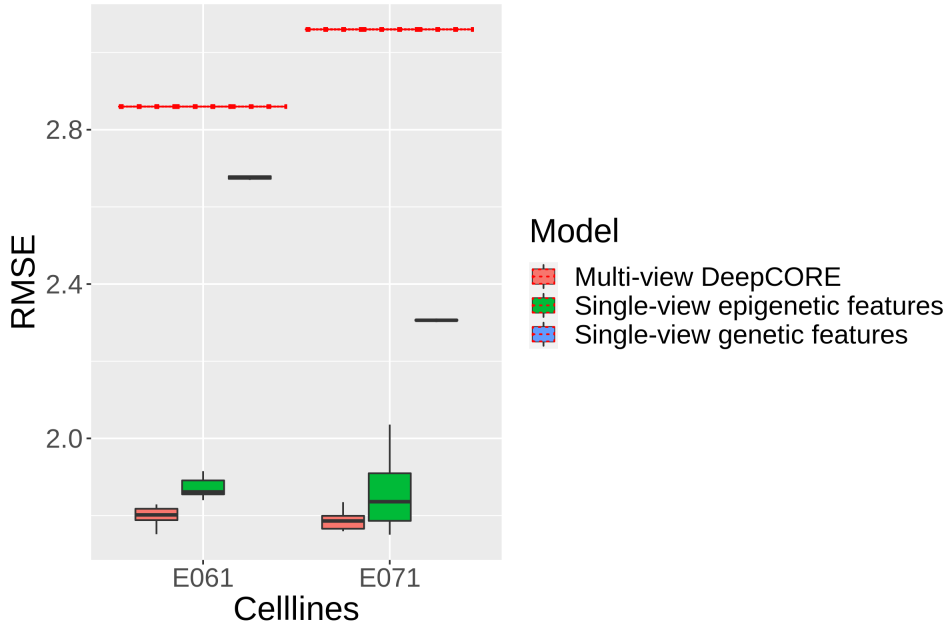


Figure 3.7: Regression Error Rate Comparison of Various Models on E061 and E071 Cell Lines. The Dashed Red Line Represents the Error Rate of ExPecto.

lower error rates ($RMSE_{E061}$=1.756, 1.789, 2.884 and $RMSE_{E071}$=1.624, 1.76, 2.393, respectively, **Figure 3.7**). We used root mean squared error as the performance measure for our analysis as our task is regression-based. An interesting point to note here is the single-view epigenetic features model performs very well and has similar

F-1scores as the multi-view model (0.834 vs. 0.845 on E061 cell line and 0.734 vs. 0.75 on E071 cell line, **Figure 3.8**). Although this might mean that genetic features might contain redundant information as epigenetic features, it still provides extra information to provide a better feature space that can describe the variations in gene expressions.

The performance of the models can be affected by the randomness in the weight initialization incurred in the deep learning model. To counter the randomness, we performed bootstrapping on the two cell lines where we randomly split the data into 20 different sets of training, testing, and validation, and used these samples to compare the performance of the models. DeepCORE performs better than single-view genetic and epigenetic feature models ($RMSE_{E061} = 1.80 \pm 0.014, 1.87 \pm 0.015, 2.68 \pm 0.004$ and $RMSE_{E071} = 1.79 \pm 0.015, 1.85 \pm 0.055, 2.31 \pm 0.002$ respectively).

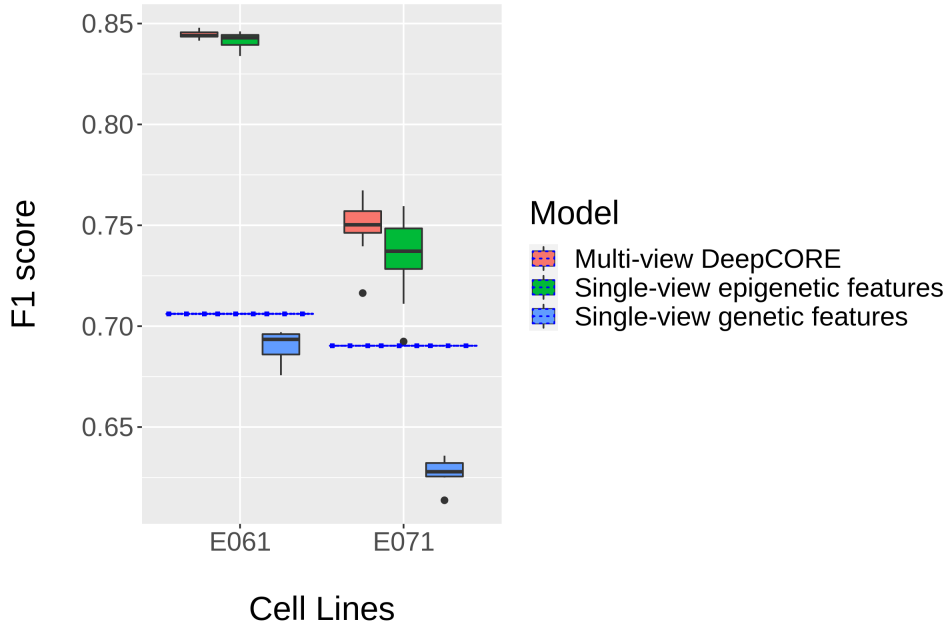


Figure 3.8: Binary Classification Performance of Various Models on E061 and E071 Cell Lines. The Dashed Blue Line Represents the F1 Score of DeepChrome.

DeepCORE can accurately predict tissue-specific gene expression level

We compared the performance of DeepCORE with ExPecto and DeepChrome (**Table 3.2**). As our DeepChrome is a binary classification model and our model was built on the predicting the continuous gene expression levels (regression), we used the median gene expression as the cutoff to convert our predictions into binary outcome to compare the two models. We used F1 score as the performance measure for the binary classification. We can see that our multi-view model outperforms both ExPecto and DeepChrome in both the cell. An interesting point to note here is the fact that the features containing epigenetic information alone perform very well and is close to the multi-view model. Although this shows that sequential information contains redundant information as epigenetics, it still provides extra information to provide a better feature space that can describe the variations in gene expressions.

Table 3.2: Performance Comparison of Various Models on E061 and E071 Cell Lines.

Cell line	**Model**	**RMSE**	**F1=-score**
E061	Single-view genetic features	2.68 ± 0.004	0.69 ± 0.37
	Single-view epigenetic features	1.87 ± 0.015	0.834 ± 0.25
	DeepCORE (multi-view)	$\mathbf{1.80 \pm 0.014}$	$\mathbf{0.845 \pm 0.11}$
	DeepChrome	NA	0.706
	ExPecto	2.86	0.821
E061	Single-view genetic features	2.31 ± 0.002	0.627 ± 0.62
	Single-view epigenetic features	1.85 ± 0.055	0.734 ± 0.84
	DeepCORE (multi-view)	$\mathbf{1.79 \pm 0.015}$	$\mathbf{0.749 \pm 0.89}$
	DeepChrome	NA	0.69
	ExPecto	3.06	**0.753**

We then predicted gene expression levels for 26 different cell lines using Deep-

CORE and ExPecto. **Figure 3.9** compares error rates (RMSE) between DeepCORE and ExPecto across 25 different cell and tissue types. It is evident that DeepCORE consistently has a lower error rate across all the cell lines over ExPecto. DeepCORE performs the best in E084 cell line (RMSE = 1.6523) and the lowest performance is achieved in E006 cell line (RMSE=2.0567).

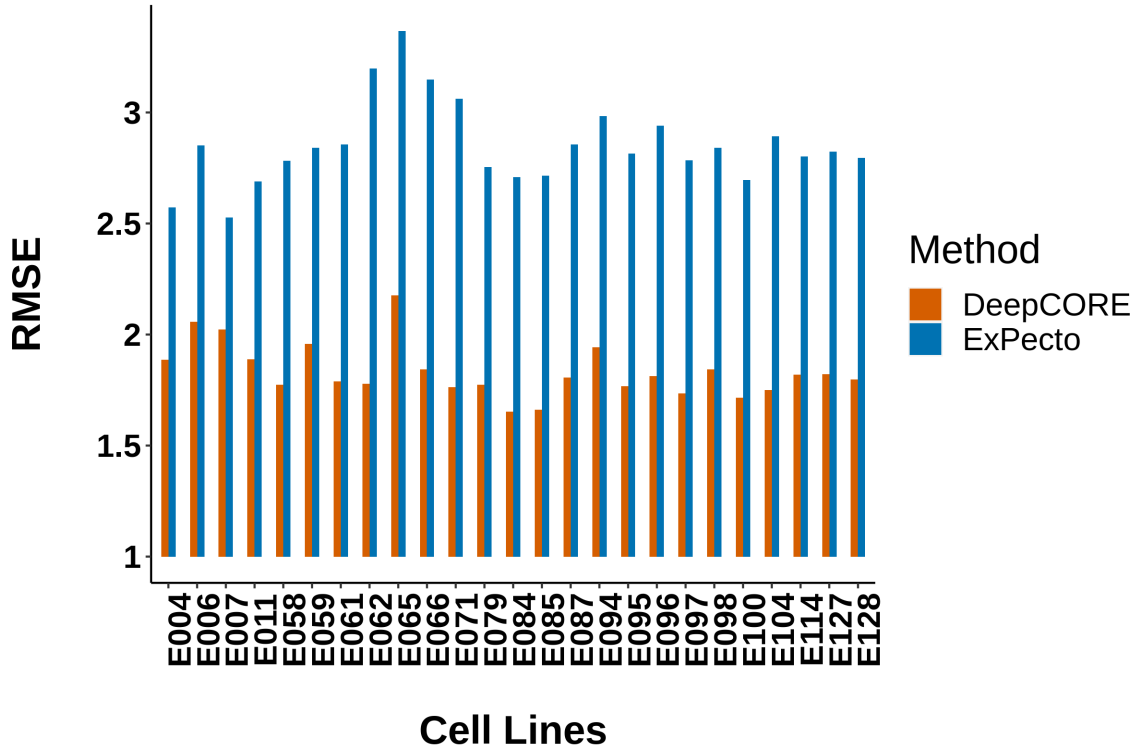


Figure 3.9: Error Rate Comparison Between DeepCORE and ExPecto Across 25 Different Cell Lines.

Similarly, we also compared DeepCORE with DeepChrome on the binary performance which is summarized in Figure 3.10. Our model has higher accuracies compared to DeepChrome across all cell lines with an improvement of more than 10% in many cell lines. The maximum accuracy was achieved by DeepChrome in E084 cell line (f1 = 78.43) and the lowest performance is on E064 cell line (f1 = 65.19).

Based on these results, we can arrive at two important deductions. Firstly, our

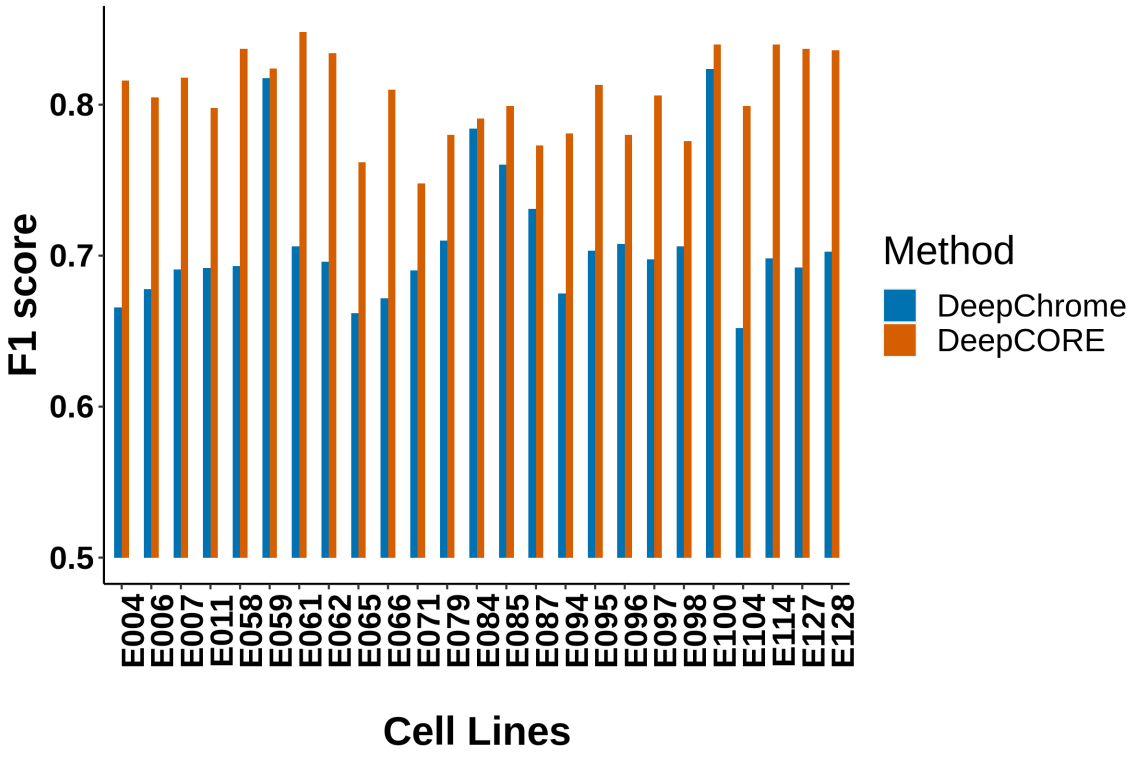


Figure 3.10: Performance Comparison Between DeepCORE and DeepChrome Across 25 Different Cell Lines.

single-view epigenetic feature model performs better than DeepChrome ($F1_{E061}$= 0.834 vs. 0.706 and $F1_{E071}$= 0.734 vs. 0.69, respectively) implying the binning strategy used by DeepChrome and many existing methods do not capture all the required features for gene expression prediction. Secondly, even though ExPecto covers four times as much region around the TSS in comparison to our model, we perform better than implying combining both genetic and epigenetic features captures more informative features than just using the sequences alone.

DeepCORE accurately predicts cross-tissue gene expression levels

We performed cross-tissue prediction to assess the generalizability of a model. For cross-tissue gene expression prediction, we used all the data from one cell line to

train the model and used it to predict the gene expression levels in another tissue. For this analysis, we used four tissues (E058, E061, E071, and E100) where we trained the model on one tissue and used this model to predict all other tissues. **Figures 3.11 and 3.12** compares the performance of DeepCORE against DeepChrome. DeepCORE preforms consistently similar to the within-tissue predictions. However, DeepChrome has huge variations in their performance with a performance drop of over 20% (F1(E100 on E071) = .583 vs F1(within) = 82.37).

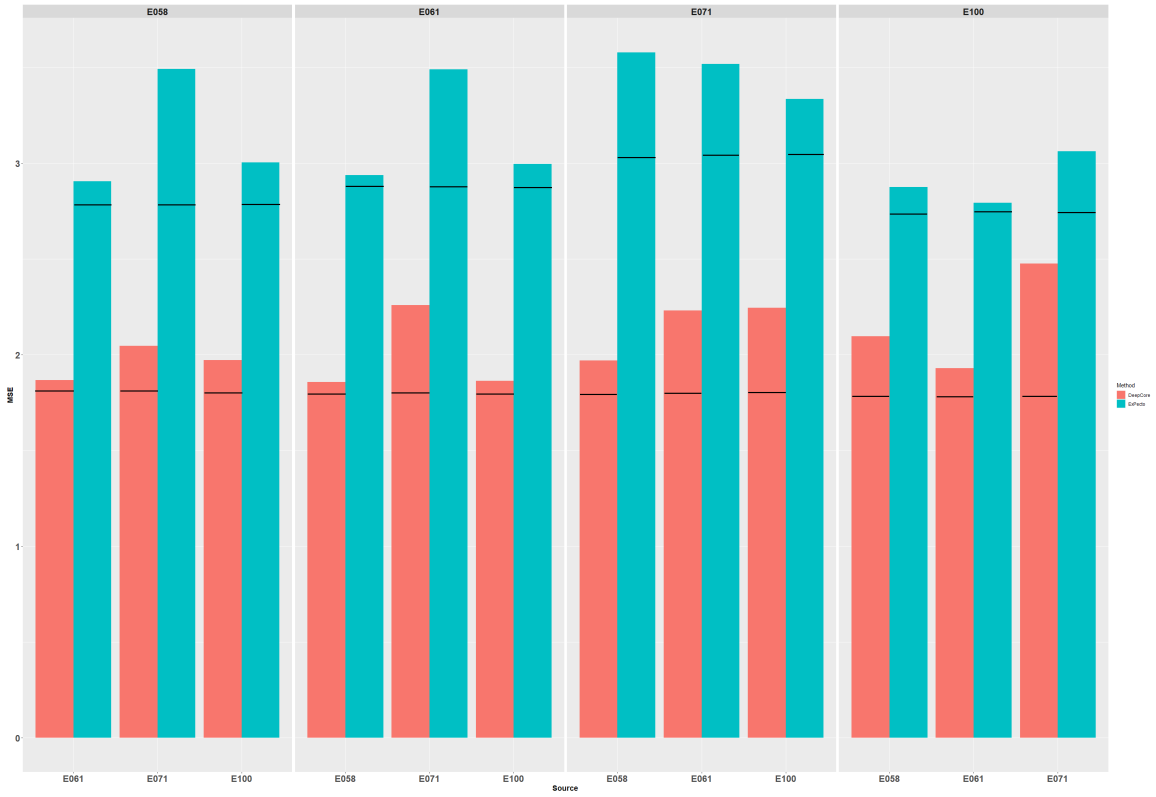


Figure 3.11: Cross-tissue Performance Comparison Between DeepCORE and DeepChrome Across 4 Cell Lines. F1 Score was Used as the Performance Metric. Black Lines Denote the Within Cell Line Accuracy.

Similar results are seen when comparing DeepChrome with ExPecto. In general, our model performs accurately irrespective of the task being within- or cross- cell line.

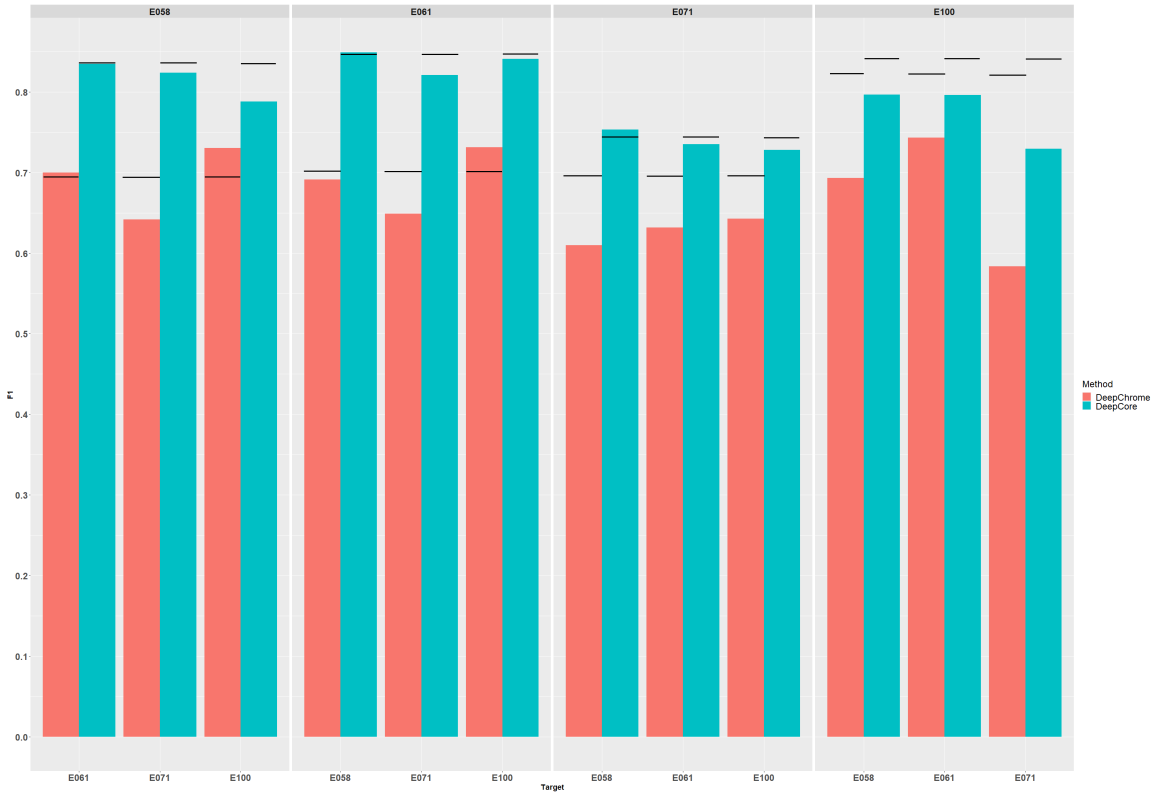


Figure 3.12: Cross-tissue Performance Comparison Between DeepCORE and ExPecto Across 4 Cell Lines. MSE is Used as the Performance Metric. Black Lines Denote the Within Cell Line Error Rate.

DeepCORE pays attention to biologically meaningful histone marks

We further investigated the biological implications of attended regions by randomly selecting 25 genes from the high expression genes and 25 low expression genes (based on the median cutoff) in E071 cell line. We extracted the attended bins that are highly significant and checked the histone mark signals in these regions.

Table 3.3 gives a count of histone marks enriched in the attended regions detected from our method. DeepCORE can not only identify regions with high histone modification activity but also detect the proper histone for active, poised, and repressed enhancers and promoters. We can see that the high expression genes have domination of H3K4me3 and H3K27ac. Active enhancer and promoters are characterized

Table 3.3: Count of Histone Marks in the Attended Regions of High and Low Expression Genes.

Histone Marks	**Low Expression Genes**		**High Expression Genes**	
	Enhancers	Promtoers	Enhancers	Promtoers
H3K4me1	1	0	4	7
H3K4me3	5	15	28	33
H3K9me3	17	34	0	0
H3K27me3	18	36	5	8
H3K4me3	1	1	42	43

by H3K27ac and H3K4me3 which regulates the gene expression represented and are concordant with the biological mechanisms of H3K27ac modifications. Similarly, we can see that the low expressed genes have H3K9me3 and H3K27me3. We know that H3K9me3 is known for down-regulation of gene and are present in poised or repressed enhancers along with H3K27me3.

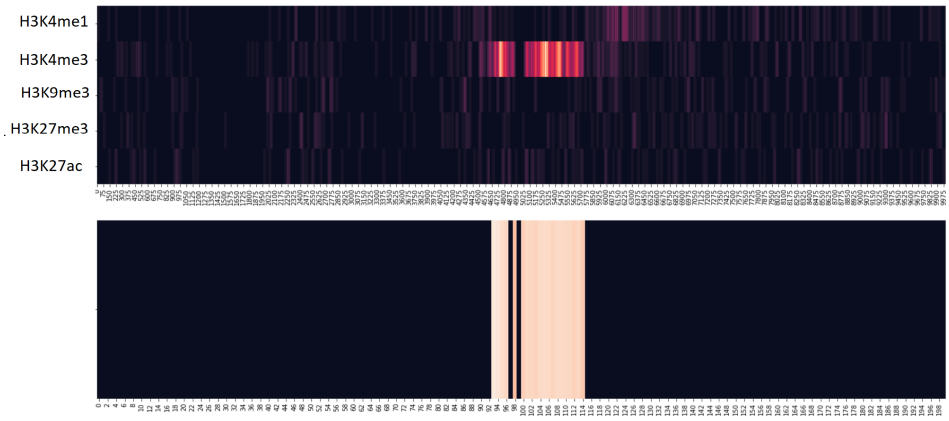


Figure 3.13: CYFIP2 Gene in E007 Cell Line. (Top) Raw Histone Modification Read Count Signals. (Bottom) DeepCORE Attended Reegions.

Figures 3.13 and 3.14 compare the attended regions of two different cell lines for CYFIP2 gene where it is highly expressed in E007 and has a very low expression

level in E058. For E007 cell line, DeepCORE pays attention only to the regions having high signals in H3K4me3 and H3K27ac (**Figure 3.13**). These two histone modifications are known to enhance gene regulation.

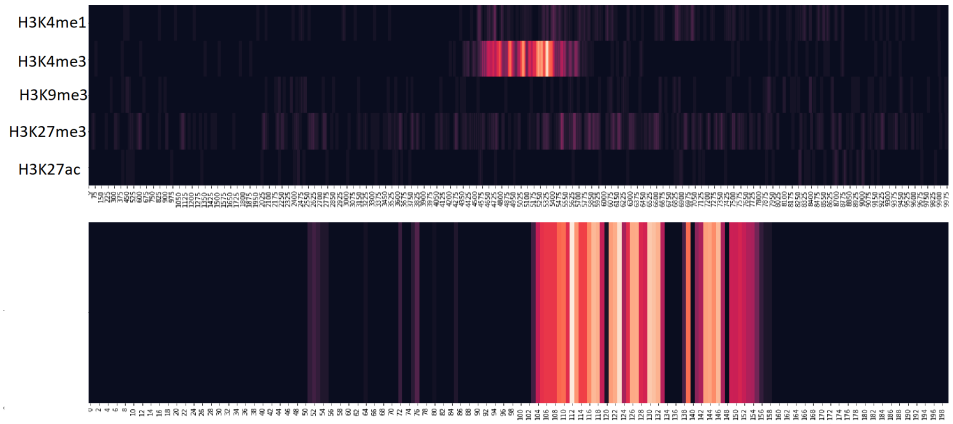


Figure 3.14: CYFIP2 Gene in E058 Cell Line. (Top) Raw Histone Modification Read Count Signals. (Bottom) DeepCORE Attended Regions.

In contrast, CYFIP2 gene is lowly expressed in E058 cell line. Our model pays attention to the repressor histone marks (H3K9me3 and H3K27me3). Along with the repressor histone signals, signals are high for H3K4me3 histone around the TSS of the gene. However, our model doesn't pay attention to these regions (**Figures 3.14**).

DeepCORE can accurately identify regulatory elements

To test the effectiveness of our model in detecting COREs, we first examined how our attention is performing in terms of detecting true histone marks. For this experiment, we compared the counts of attended bins with histone marks and attended bins without histone marks for a randomly selected 155 genes across all the cell lines. **Figure 3.15** shows the density plot of the comparison. In most of the cases, all attended bins contain some level of histone variation signals (ttest p-value = 0.0). We removed all the attended bins with no histone signals for enrichment analysis of regulatory elements in these attended regions.

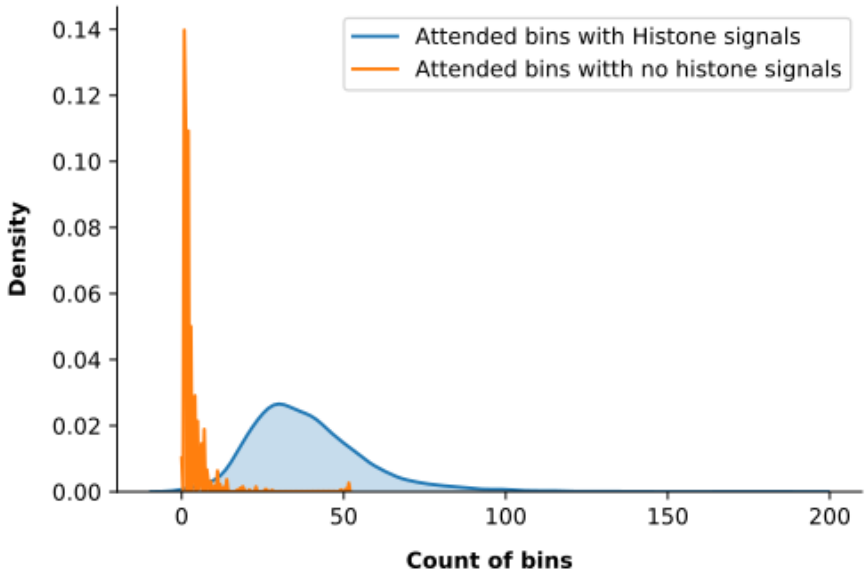
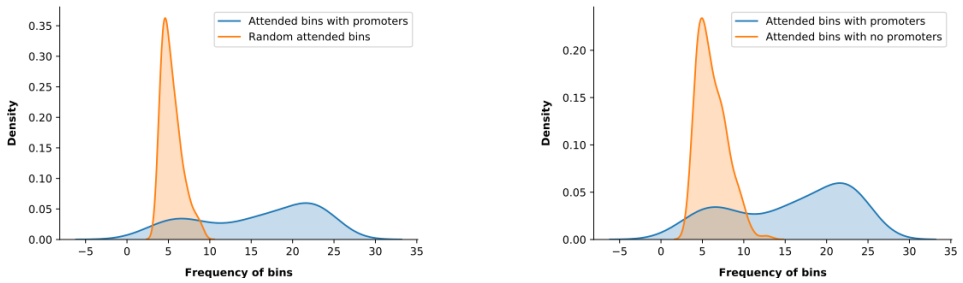


Figure 3.15: Density Plot of the Count of Attended Bins with Histone Signals Vs. Count of Attended Bins with No Histone Signals.

Eukaryotic Promoter Database (EPD, Périer *et al.* (2000)) contains a comprehensive list of experimentally validated eukaryotic promoters for various organisms. We extracted 29598 human promoters from EPD. Similarly, we also extracted 250,512 experimentally validated human candidate enhancer from GeneHancer (Fishilevich *et al.*, 2017). We then scanned our attended regions to identify the presence of promoters or enhancers. For matching the attended regions with the promoters, we restricted the attended regions to be within 1000 bps around the TSS. No such restrictions were applied for matching enhancers.

We hypothesized that known regulatory elements are ranked higher in our attention matrix. To test our hypothesis, we calculated the frequency of the attended regions containing known regulatory elements across all cell lines and the frequency of the remaining regions. We first compared the differences in the average frequencies of attended bins with known promoters against random attended bins with some

histone signals (**Figure 3.16a**).

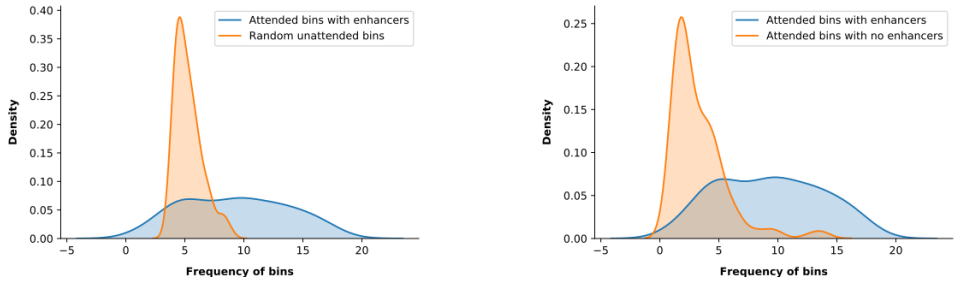


(a) (b)

Figure 3.16: Density Plot of the Frequency of Attended Bins with Known Promoters Across 25 Cell Lines. (a) Attended Bins with Promoters Vs. Random Attended Bins (B) Attended Bins with Promoters Vs. Attended Bins Without Promoters.

The distribution for randomly chosen attended bins with histone marks is very narrow with a mean frequency of 5 whereas the spread in the frequency of attended bins with promoters is very wide with two peaks (frequency of 5 and frequency of 25). The difference between the two groups is highly significant (Mann Whitney p-value = $4.002e^{-25}$). **Figure 3.16b** plots the differences in the average frequencies of attended bins with promoters and bins without promoters. The distribution is narrow for bins without promoters with a peak at around frequency of 5. On the contrary, the distribution is wider for attended regions with promoters with a peak at around frequency of 25 which is a fivefold increase over the frequency of bins attended with no promoters (Mann Whitney p-value = $9.118e^{-25}$).

We see similar results for the attended bins with enhancer regions. **Figure 3.17a** compares frequency of attended bins with known enhancers against choosing random attended bins with no histone modification signals whereas **Figure 3.17b** compares frequency of attended bins with known enhancers against attended bins without enhancers. We see similar results for the enhancer regions as well (Mann Whitney test: p-values = $6.116e^{-18}$ and $7.3933e^{-18}$ respectively). However, the gap between the av-



(a) (b)

Figure 3.17: Density Plot of the Frequency of Attended Bins with Known Enhancers Across 25 Cell Lines. (a) Attended Bins with Enhancers Vs. Random Unattended Bins (B) Attended Bins with Enhancers Vs. Attended Bins Without Enhancers.

erage frequency of attended bins without enhancers and with enhancers is narrower (mean frequency of 3 vs. peak at two different locations (mean frequency of 5 and mean frequency of 10) for attended bins with enhancer elements). This shows that our model indeed pays attention to the genomic regions that are enriched with known promoters and enhancers.

Figure 3.18 shows the regulatory elements detection in TRAK2 gene in E006 cell line. This gene is highly expressed which is associated with high histone signals in H3K4me3 around the TSS of the gene. DeepCORE attends to these high activator signals that are also enriched with known enhancers and promoters.

Figure 3.19 also shows how attended bins indeed are enriched with promoters and enhancers. TMEM88 gene is highly expressed in E004 cell line which is characterized by H3K4me1, H3K4me3, and H3K27ac histone modification signals. EPD promoters for this gene lie in three different regions, one right before the transcription start site and the other two towards the downstream of the gene. The enhancers span a wide range starting at 1200 bps upstream of the TSS till 4000 bps downstream of the TSS of the gene. We can see that even though there are high signals for the H3K27me3 which act as repressors, our model does not pay attention to these regions, and also

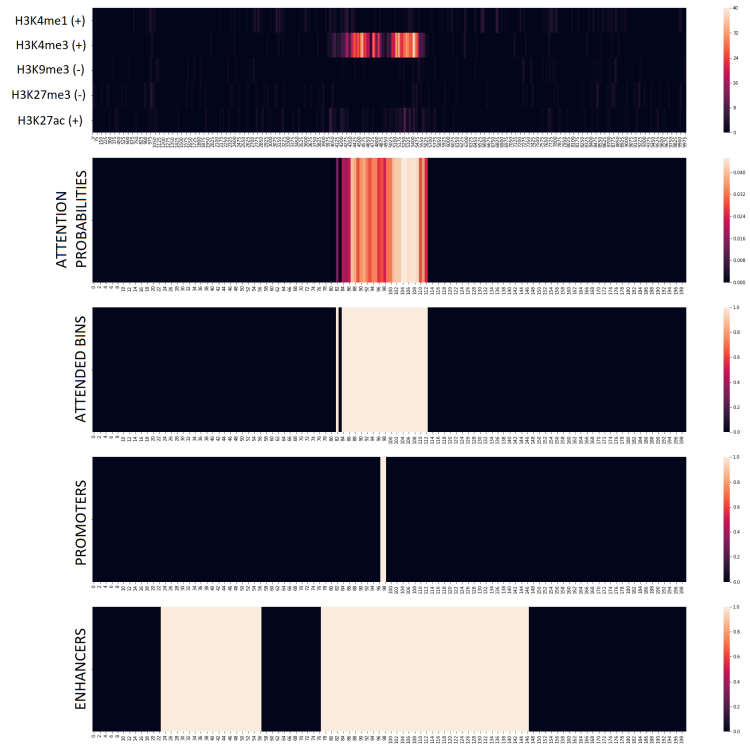


Figure 3.18: Mapping Attended Bins With Known Promoters and Enhancers for TRAK2 Gene in E006 Cell Line.

these regions are not enriched with any promoters and enhancers.

DeepCORE performs fine mapping of COREs

Figure 3.20 compares the attention picked up by our model in three different cell lines (E095, E065, and E100). The raw histone signals have two different peaks in all the three cell lines (the first peak surrounding the TSS and the second one is around 2500 bps upstream of the TSS).

On close observation of this gene in the UCSC genome browser, the left peak (2.5 Kbps upstream) belongs to another gene (GCC1) and the REMC annotates both these peaks belonging to ARFS5 as both these genes have some overlapping regions and are co-expressed. However, our model correctly identifies the histone signals

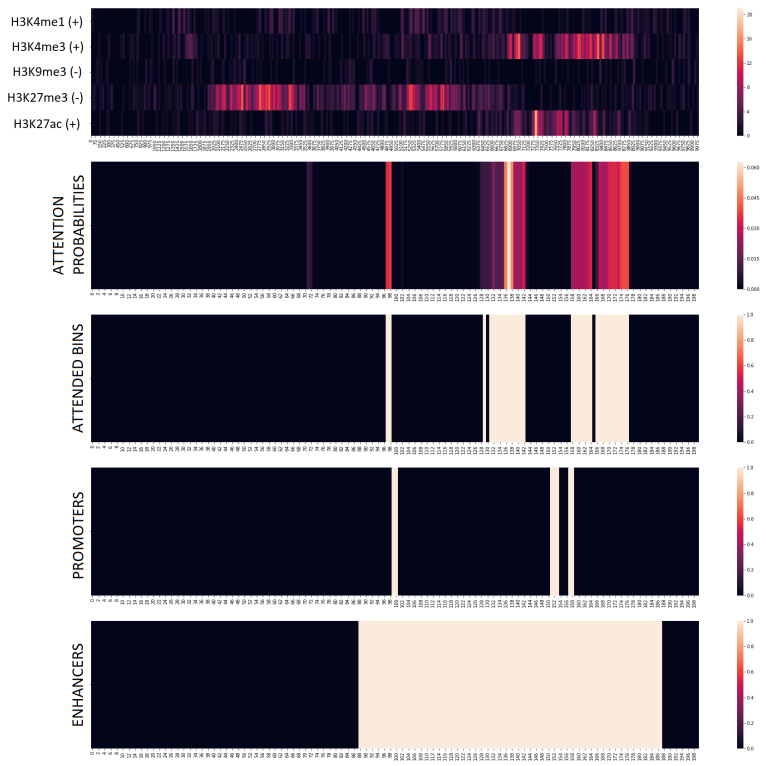


Figure 3.19: Mapping Attended Bins with Known Promoters and Enhancers for TMEM88 Gene in E004 Cell Line.

corresponding to ARF5 gene only across all the tissues. We do pick up some signals from the left peak for E095 cell line, but the intensity of the attention score for this region is very low (attention probability for the left peak is 0.006 vs. right peak being 0.03).

GeneHancer has a wider spread of enhancer regions (almost spanning half of the genomic regions within our data, **Figures 3.18 and 3.19**). This can mean that there can be multiple enhancer regions that can bind to the promoters and picking the right region is a challenge from just the annotations. However, the attended bins picked by our model have a higher resolution with narrow genomic regions that are the true contributors to gene expression prediction meaning DeepCORE performs fine mapping of functional regulatory elements.

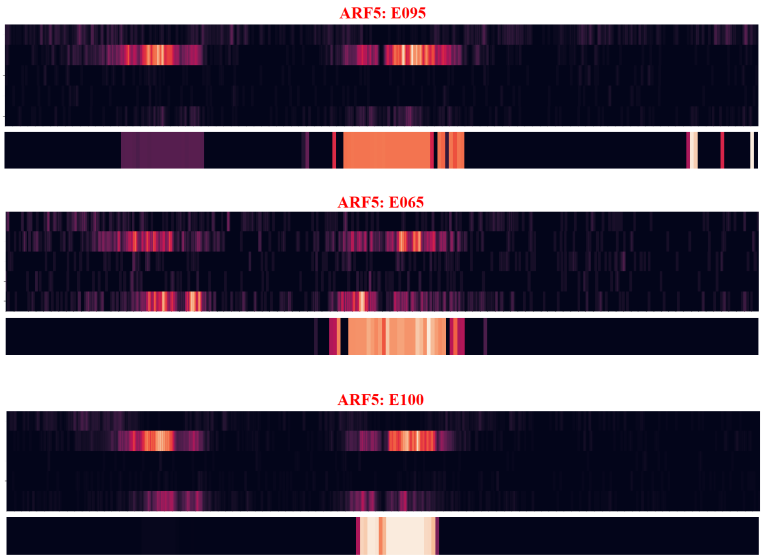


Figure 3.20: Fine Mapping of Genomic Regions in ARF5 Gene Across 3 Different Cell Lines. For Each Cell Line, the Top Portion Denotes Histone Signals and the Bottom Portion Represents the Attention.

To further investigate the fine mapping of COREs, we performed gene-specific detection of co-operative regulatory elements that are consistently activated across all cell lines contributing towards gene regulation. To do this, we performed pair-wise correlation on the attended regions across all the cell lines and filtered all the pairs which had a low value based on the correlation threshold of 0.9 and a p-value threshold of 0.05. To account for multiple comparison corrections, we used Bonferroni correction technique. Figure 3.20 gives the correlation plot of PSMD8 gene after filtration.

PSMD8 gene is consistently highly expressed across 25 different cell lines. The highly significant attended bins across all the cell lines lie around the TSS of this gene ranging from 1000 bps upstream to 2000 bps downstream. Our model picks up two highly correlated regions (A and B in **Figure 3.21**). The first region is a highly dense region and the second one is a sparse region along with sparse pairs of highly correlated bins. The sparsity helps us to pinpoint the exact genomic regions that are functional in regulating the gene expression. This shows that our model provides finer

resolution in terms of detecting cooperative functional regulatory elements which is in other term known as fine mapping. We believe cataloging these co-operative functional regulatory non-coding regions is the key to bridging the gap in understanding genotype-phenotype association.

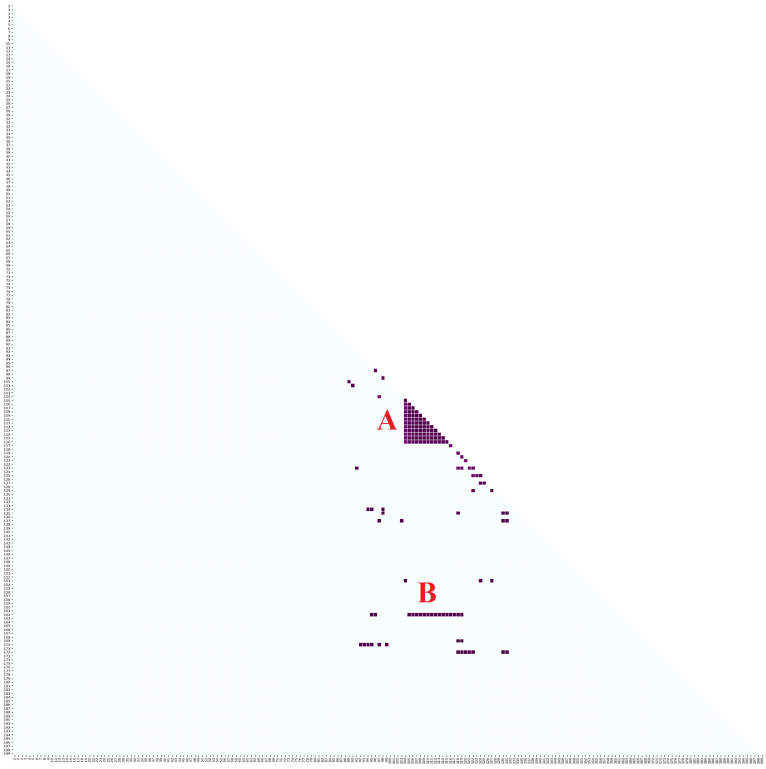


Figure 3.21: Correlation Plot on PSMD8 Gene.

3.4 Limitations and Future Direction

We foresee several ways to further improve the DeepCORE methods. Firstly, we assume that only a limited set of histone modifications and transcription binding sites affect gene expression level due to the lack of data availability and only one type of epigenetic markers in terms of histone modification signals are influencing gene expressions. Biologically, the gene regulatory mechanism is more complex than this assumption and having more data can improve the prediction power and COREs

detection. Secondly, other epigenetic factors such as DNA methylation data provide complementary information to histone modifications. Including these marks in the neural network model may increase the prediction accuracy and enhance the attention matrices. Secondly, expanding the genomic regions from the current 10Kbps to 2Mbps flanking TSSs may allow us to discover distal regulatory elements. Thirdly, integrating multiple tissues in the same neural network may increase the resolution of DeepCORE. Currently, the predicted COREs span several hundreds of bps. Because enhancers are often clustered, and activation of different enhancers in the same cluster is tissue-specific (Pennacchio *et al.*, 2007; Ong and Corces, 2011; Xiong *et al.*, 2018), concurrent modeling of multiple tissues is promising to capture the boundary between these enhancers and subsequently increase the resolution. This will also identify tissue-specific gene-promoter and gene-enhancer interactions, which is extremely valuable knowledge that has not been annotated in existing databases.

Finally, our model does not account for the impact of genetic signals on epigenetic variations (an arrow between genetic signals and epigenetic signals in **Figure 3.6**). As there is a lack of data availability in this aspect, incorporating this genetic to epigenetic flow is in our future interests.

Chapter 4

CONCLUSION AND FUTURE WORK

In Chapter 1, we reviewed the state-of-the-art computational approaches to delineating functional genomes and variants in noncoding regions. Via extensive simulations, we reported significant knowledge gaps in identifying regulatory elements and variants in noncoding genomes. Firstly, the resolution of computationally predicted functional elements is insufficient to distinguish functionally neutral vs. non-neutral variants in close vicinity, or neutral vs. non-neutral alleles at the same position Secondly, the lack of consideration of epistasis between regulatory elements makes current methods powerless to discover regulatory elements or variants that function collaboratively. To address these knowledge gaps, we have developed two novel methods using advanced machine-learning approaches.

In Chapter 2, we presented the first method, TreepMap that fine maps regulatory variants across multiple sites. TreeMap contains three layers that assess the regulatory associations of genetic elements at increasing resolutions, from coarse-grained LD blocks to fine-grained positions, to single nucleotides. At each layer, TreeMap allows multi-site effects. We show that TreeMap detects co-regulating genetic variants accurately when compared with the current approaches. The application of TreeMap to GTEx brain and colon datasets reported pervasive epistasis among regulatory variants and discovered hundreds of novel eQTL loci.

In Chapter 3, we presented another novel method, DeeCORE that identifies cooperative regulatory elements (COREs) by modeling the relationship between genetic, epigenetic and transcriptional patterns across tissues and cell lines. DeepCORE is built upon a novel attention-based multi-view deep neural network architecture. This

neural network is trained to predict tissue-specific gene expression based on genetic sequences and histone modifications, which significantly outperforms the existing approaches. Downstream analysis of the attentions embedded in this neural network informed locations of regulatory elements and their correlations, which collectively implies COREs. We also show these COREs are significantly enriched with promoters and enhancers.

4.1 Integrating TreeMap and DeepCORE

Our future work includes combining the two proposed algorithms, TreeMap and DeepCORE, to improve fine mapping of functional elements (variants and regulatory elements). As we have seen from Figure 2.4a, adding functional annotations can significantly improve the prediction performance of TreeMap. Attentions derived from DeepCORE provide functional information of noncoding elements. Therefore, it is natural to use DeepCORE attentions as functional scores in TreeMap. Because DeepCORE attentions are tissue-specific and have higher resolution than raw histone marks or existing promoter/enhancer annotations, we expect merging DeepCORE with TreeMap will enable accurate fine mapping informed with tissue-specific functional knowledge.

4.2 Extending DeepCORE for EWAS

GWAS studies have proven to be a huge success in identifying the genotype loci associated with phenotypes. However, there is a need to understand the epigenetic impact on the phenotype as epigenetic marks are altered throughout the life modified by various environmental factors (Rakyan et al., 2011; Verma, 2012; Birney et al., 2016). The consequences of these alterations are reflected in terms of changes in the gene expression level thereby causing diseases. In line with this, EWAS studies

have just started to be more mainstream and a lot of EWAS is focused on DNA methylation. The EWAS studies aim at mapping the association between epigenetic perturbations to phenotype changes. Understanding these EWAS studies has vast therapeutic potential. For example, oridonin, a chemical compound, was discovered to be effective against colorectal cancer where this compound played a major role in the activation of various histone marks and suppression of few genes (Gao et al., 2010). An ideal scenario is where we can combine the EWAS and GWAS results in understanding the overall mechanism underlying the phenotypic variations.

From this point of view, the proposed DeepCORE architecture can be extended to EWAS studies. By altering the genetic and epigenetic signals to mimic genetic variants and epigenetic changes between healthy and disease groups, we can assess their potential transcriptional impact. Significant transcriptional impact estimated in these in silico perturbations can inform the evaluation of pathogenic variants and potential disease mechanisms. This knowledge will provide valuable guidance to further investigation using experimental approaches.

REFERENCES

Adzhubei, I., D. M. Jordan and S. R. Sunyaev, "Predicting functional effect of human missense mutations using polyphen-2", Current protocols in human genetics **76**, 1, 7–20 (2013).

Alipanahi, B., A. Delong, M. T. Weirauch and B. J. Frey, "Predicting the sequence specificities of dna-and rna-binding proteins by deep learning", Nature biotechnology **33**, 8, 831–838 (2015).

Angermueller, C., T. Pärnamaa, L. Parts and O. Stegle, "Deep learning for computational biology", Molecular systems biology **12**, 7, 878 (2016).

Appella, E. and C. W. Anderson, "Post-translational modifications and activation of p53 by genotoxic stresses", European Journal of Biochemistry **268**, 10, 2764–2772 (2001).

Bahdanau, D., K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate", arXiv preprint arXiv:1409.0473 (2014).

Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada *et al.*, "Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease", Nature genetics **40**, 8, 955–962 (2008).

Barutcuoglu, Z., R. E. Schapire and O. G. Troyanskaya, "Hierarchical multi-label prediction of gene function", Bioinformatics **22**, 7, 830–836 (2006).

Beer, M. A. and S. Tavazoie, "Predicting gene expression from sequence", Cell **117**, 2, 185–198 (2004).

Benesty, J., J. Chen, Y. Huang and I. Cohen, "Pearson correlation coefficient", in "Noise reduction in speech processing", pp. 1–4 (Springer, 2009).

Bengio, Y., P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult", IEEE transactions on neural networks **5**, 2, 157–166 (1994).

Benner, C., C. C. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti and M. Pirinen, "Finemap: efficient variable selection using summary data from genome-wide association studies", Bioinformatics **32**, 10, 1493–1501 (2016).

Bernstein, B. E., J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker *et al.*, "The nih roadmap epigenomics mapping consortium", Nature biotechnology **28**, 10, 1045–1048 (2010).

Bhalala, O. G., A. P. Nath, M. Inouye, C. R. Sibley, U. B. E. Consortium *et al.*, "Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue", PLoS genetics **14**, 8, e1007607 (2018).

Blackwood, E. M. and J. T. Kadonaga, "Going the distance: a current view of enhancer action", Science **281**, 5373, 60–63 (1998).

Boloor, A., X. He, C. Gill, Y. Vorobeychik and X. Zhang, "Simple physical adversarial examples against end-to-end autonomous driving models", in "2019 IEEE International Conference on Embedded Software and Systems (ICESS)", pp. 1–7 (IEEE, 2019).

Brāzma, A., I. Jonassen, J. Vilo and E. Ukkonen, "Predicting gene regulatory elements in silico on a genomic scale", Genome research **8**, 11, 1202–1215 (1998).

Brown, A. A., A. Viñuela, O. Delaneau, T. D. Spector, K. S. Small and E. T. Dermitzakis, "Predicting causal variants affecting expression by using whole-genome sequencing and rna-seq from multiple human tissues", Nature genetics **49**, 12, 1747–1751 (2017).

Buske, O. J., A. Manickaraj, S. Mital, P. N. Ray and M. Brudno, "Identification of deleterious synonymous variants in human genomes", Bioinformatics **29**, 15, 1843–1850 (2013).

Bussemaker, H. J., H. Li and E. D. Siggia, "Regulatory element detection using correlation with expression", Nature genetics **27**, 2, 167–171 (2001).

Cai, J. J., J. M. Macpherson, G. Sella and D. A. Petrov, "Pervasive hitchhiking at coding and regulatory sites in humans", PLoS genetics **5**, 1, e1000336 (2009).

Cannon, M. E. and K. L. Mohlke, "Deciphering the emerging complexities of molecular mechanisms at gwas loci", The American Journal of Human Genetics **103**, 5, 637–653 (2018).

Cao, S. and R. Nevatia, "Exploring deep learning based solutions in fine grained activity recognition in the wild", in "2016 23rd International Conference on Pattern Recognition (ICPR)", pp. 384–389 (IEEE, 2016).

Carey, M., C. L. Peterson and S. T. Smale, *Transcriptional regulation in eukaryotes* (Cold Spring Harbor Laboratory Press, 2009).

Chatterjee, S. and N. Ahituv, "Gene regulatory elements, major drivers of human disease", Annual review of genomics and human genetics **18**, 45–63 (2017).

Chen, C., A. Seff, A. Kornhauser and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving", in "Proceedings of the IEEE International Conference on Computer Vision", pp. 2722–2730 (2015a).

Chen, L., "Attention-based deep learning system for negation and assertion detection in clinical notes", International Journal of Artificial Intelligence and Applications (IJAIA) **10**, 1 (2019).

Chen, L. and J. Widom, "Mechanism of transcriptional silencing in yeast", Cell **120**, 1, 37–48 (2005).

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation", in "Proceedings of the European conference on computer vision (ECCV)", pp. 801–818 (2018).

Chen, Q.-Y., T. Zhang, S. H. Pincus, S. Wu, D. Ricks, D. Liu, Z. Sun, N. Maclaren and M. S. Lan, "Human cd1d gene expression is regulated by lef-1 through distal promoter regulatory elements", The Journal of Immunology **184**, 9, 5047–5054 (2010).

Chen, W., B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, G. A. Poland and D. J. Schaid, "Fine mapping causal variants with an approximate bayesian method using marginal test statistics", Genetics **200**, 3, 719–736 (2015b).

Chen, Y., Y. Li, R. Narayan, A. Subramanian and X. Xie, "Gene expression inference with deep learning", Bioinformatics **32**, 12, 1832–1839 (2016).

Cheng, C. and M. Gerstein, "Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells", Nucleic acids research **40**, 2, 553–568 (2012).

Cheng, C., K.-K. Yan, K. Y. Yip, J. Rozowsky, R. Alexander, C. Shou and M. Gerstein, "A statistical framework for modeling gene expression using chromatin features and application to modencode datasets", Genome biology **12**, 2, 1–18 (2011).

Chiu, J. P. and E. Nichols, "Named entity recognition with bidirectional lstm-cnns", Transactions of the Association for Computational Linguistics **4**, 357–370 (2016).

Cho, K., B. Van Merriënboer, D. Bahdanau and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches", arXiv preprint arXiv:1409.1259 (2014a).

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation", arXiv preprint arXiv:1406.1078 (2014b).

Ciompi, F., B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop and B. van Ginneken, "Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box", Medical image analysis **26**, 1, 195–202 (2015).

Clauset, A., M. E. Newman and C. Moore, "Finding community structure in very large networks", Physical review E **70**, 6, 066111 (2004).

Conlon, E. M., X. S. Liu, J. D. Lieb and J. S. Liu, "Integrating regulatory motif discovery and genome-wide expression analysis", Proceedings of the National Academy of Sciences **100**, 6, 3339–3344 (2003).

Consortium, . G. P. *et al.*, "A global reference for human genetic variation", Nature **526**, 7571, 68–74 (2015).

Consortium, E. P. *et al.*, "The encode (encyclopedia of dna elements) project", Science **306**, 5696, 636–640 (2004).

Consortium, G. *et al.*, "Genetic effects on gene expression across human tissues", Nature **550**, 7675, 204–213 (2017).

Corona, E., J. T. Dudley and A. J. Butte, "Extreme evolutionary disparities seen in positive selection across seven complex diseases", PloS one **5**, 8, e12236 (2010).

Costa, I. G., H. G. Roider, T. G. do Rego and F. d. A. de Carvalho, "Predicting gene expression in t cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models", BMC bioinformatics **12**, 1, 1–10 (2011).

Cui, Z., R. Ke, Z. Pu and Y. Wang, "Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction", arXiv preprint arXiv:1801.02143 (2018).

Curtis, D., R. Lehmann and P. D. Zamore, "Translational regulation in development", Cell **81**, 2, 171–178 (1995).

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, "High-resolution haplotype structure in the human genome", Nature genetics **29**, 2, 229–232 (2001).

Dong, L., F. Wei, M. Zhou and K. Xu, "Question answering over freebase with multi-column convolutional neural networks", in "Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)", pp. 260–269 (2015).

Dong, X., M. C. Greven, A. Kundaje, S. Djebali, J. B. Brown, C. Cheng, T. R. Gingeras, M. Gerstein, R. Guigó, E. Birney *et al.*, "Modeling gene expression using chromatin features in various cellular contexts", Genome biology **13**, 9, 1–17 (2012).

Doulamis, N., "Adaptable deep learning structures for object labeling/tracking under dynamic visual environments", Multimedia Tools and Applications **77**, 8, 9651–9689 (2018).

Doulamis, N. and A. Voulodimos, "Fast-mdl: Fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification", in "2016 IEEE International Conference on Imaging Systems and Techniques (IST)", pp. 318–323 (IEEE, 2016).

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore and J. H. Nadeau, "Missing heritability and strategies for finding the underlying causes of complex disease", Nature Reviews Genetics **11**, 6, 446–450 (2010).

Eser, U. and L. S. Churchman, "Fiddle: An integrative deep learning framework for functional genomic data inference", Biorxiv p. 081380 (2016).

Fan, Y., Y. Qian, F.-L. Xie and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks", in "Fifteenth Annual Conference of the International Speech Communication Association", (2014).

Ferreira, T. and J. Marchini, "Modeling interactions with known risk loci—a bayesian model averaging approach", Annals of human genetics **75**, 1, 1–9 (2011).

Fishilevich, S., R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran *et al.*, "Genehancer: genome-wide integration of enhancers and target genes in genecards", Database **2017** (2017).

Fourcade, A. and R. Khonsari, "Deep learning in medical image analysis: A third eye for doctors", Journal of stomatology, oral and maxillofacial surgery **120**, 4, 279–288 (2019).

Frazer, K. A., S. S. Murray, N. J. Schork and E. J. Topol, "Human genetic variation and its contribution to complex traits", Nature Reviews Genetics **10**, 4, 241–251 (2009).

Gaffney, D. J., J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad and J. K. Pritchard, "Dissecting the regulatory architecture of gene expression qtls", Genome biology **13**, 1, R7 (2012).

Gao, M., U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers *et al.*, "Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks", Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **6**, 1, 1–6 (2018).

Gelfman, S., Q. Wang, K. M. McSweeney, Z. Ren, F. La Carpia, M. Halvorsen, K. Schoch, F. Ratzon, E. L. Heinzen, M. J. Boland *et al.*, "Annotating pathogenic non-coding variants in genic regions", Nature communications **8**, 1, 1–11 (2017).

Gilad, Y., S. A. Rifkin and J. K. Pritchard, "Revealing the architecture of gene regulation: the promise of eqtl studies", Trends in genetics **24**, 8, 408–415 (2008).

Graves, A., N. Jaitly and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm", in "2013 IEEE workshop on automatic speech recognition and understanding", pp. 273–278 (IEEE, 2013).

Graves, A. and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures", Neural networks **18**, 5-6, 602–610 (2005).

Guan, X., G. Runger and L. Liu, "Dynamic incorporation of prior knowledge from multiple domains in biomarker discovery", BMC bioinformatics **21**, 2, 1–10 (2020).

Guo, Y., Y. Gao and D. Shen, "Deformable mr prostate segmentation via deep feature learning and sparse patch matching", IEEE transactions on medical imaging **35**, 4, 1077–1089 (2015).

Hinton, G., O. Vinyals and J. Dean, "Distilling the knowledge in a neural network", arXiv preprint arXiv:1503.02531 (2015).

Hobert, O., "Gene regulation by transcription factors and microrna", Science **319**, 5871, 1785–1786 (2008).

Hochreiter, S. and J. Schmidhuber, "Long short-term memory", Neural computation **9**, 8, 1735–1780 (1997).

Holzinger, A., J. Schantl, M. Schroettner, C. Seifert and K. Verspoor, "Biomedical text mining: state-of-the-art, open problems and future challenges", in "Interactive knowledge discovery and data mining in biomedical informatics", pp. 271–300 (Springer, 2014).

Hoppe, K. L. and O. L. Francone, "Binding and functional effects of transcription factors sp1 and sp3 on the proximal human lecithin: cholesterol acyltransferase promoter", Journal of lipid research **39**, 5, 969–977 (1998).

Horie, Y., T. Yoshio, K. Aoyama, S. Yoshimizu, Y. Horiuchi, A. Ishiyama, T. Hirasawa, T. Tsuchida, T. Ozawa, S. Ishihara *et al.*, "Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks", Gastrointestinal endoscopy **89**, 1, 25–32 (2019).

Hu, Z.-Z., K. B. Cohen, L. Hirschman, A. Valencia, H. Liu, M. G. Giglio and C. H. Wu, "iprolink: A framework for linking text mining with ontology and systems biology", in "2008 IEEE International Conference on Bioinformatics and Biomedicine", pp. 467–472 (IEEE, 2008).

Huang, Y.-F., B. Gulko and A. Siepel, "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data", Nature genetics **49**, 4, 618–624 (2017).

Ioannidis, N. M., J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi *et al.*, "Revel: an ensemble method for predicting the pathogenicity of rare missense variants", The American Journal of Human Genetics **99**, 4, 877–885 (2016).

Ionita-Laza, I., K. McCallum, B. Xu and J. D. Buxbaum, "A spectral approach integrating functional genomic annotations for coding and noncoding variants", Nature genetics **48**, 2, 214 (2016).

Jang, E., S. Gu and B. Poole, "Categorical reparameterization with gumbel-softmax", arXiv preprint arXiv:1611.01144 (2016).

Jeyaraj, P. R. and E. R. S. Nadar, "Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm", Journal of cancer research and clinical oncology **145**, 4, 829–837 (2019).

Jones, W., K. Alasoo, D. Fishman and L. Parts, "Computational biology: deep learning", Emerging Topics in Life Sciences **1**, 3, 257–274 (2017).

Josephs, E. B., Y. W. Lee, J. R. Stinchcombe and S. I. Wright, "Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression", Proceedings of the National Academy of Sciences **112**, 50, 15390–15395 (2015).

Keleş, S., M. van der Laan and M. B. Eisen, "Identification of regulatory elements using a feature selection method", Bioinformatics **18**, 9, 1167–1175 (2002).

Khoury, G. and P. Gruss, "Enhancer elements", Cell **33**, 2, 313–314 (1983).

Kichaev, G., W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft and B. Pasaniuc, "Integrating functional data to prioritize causal variants in statistical fine-mapping studies", PLoS Genet **10**, 10, e1004722 (2014).

Kim, H.-Y., "Statistical notes for clinical researchers: post-hoc multiple comparisons", Restorative dentistry & endodontics **40**, 2, 172–176 (2015).

Kim, J., V. D. Calhoun, E. Shim and J.-H. Lee, "Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia", Neuroimage **124**, 127–146 (2016).

Kim, Y., K. Xia, R. Tao, P. Giusti-Rodriguez, V. Vladimirov, E. Van Den Oord and P. Sullivan, "A meta-analysis of gene expression quantitative trait loci in brain", Translational psychiatry **4**, 10, e459–e459 (2014).

Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980 (2014).

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants", Nature genetics **46**, 3, 310–315 (2014).

Kleesiek, J., G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus and A. Biller, "Deep mri brain extraction: A 3d convolutional neural network for skull stripping", NeuroImage **129**, 460–469 (2016).

Koehn, P. and R. Knowles, "Six challenges for neural machine translation", in "Proceedings of the First Workshop on Neural Machine Translation", pp. 28–39 (Association for Computational Linguistics, Vancouver, 2017), URL `https://www.aclweb.org/anthology/W17-3204`.

Kouzarides, T., "Chromatin modifications and their function", Cell **128**, 4, 693–705 (2007).

Krivega, I. and A. Dean, "Enhancer and promoter interactions—long distance calls", Current opinion in genetics & development **22**, 2, 79–85 (2012).

Krizhevsky, A., I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Communications of the ACM **60**, 6, 84–90 (2017).

Kulmanov, M., M. A. Khan and R. Hoehndorf, "Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier", Bioinformatics **34**, 4, 660–668 (2018).

Kumar, S., M. Sanderford, V. E. Gray, J. Ye and L. Liu, "Evolutionary diagnosis method for variants in personal exomes", Nature methods **9**, 9, 855–856 (2012).

Kundaje, A., W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller *et al.*, "Integrative analysis of 111 reference human epigenomes", Nature **518**, 7539, 317–330 (2015).

Lamúrias, A. F. M., "Development of a text mining approach to disease network discovery", (2019).

Lanchantin, J., R. Singh, Z. Lin and Y. Qi, "Deep motif: Visualizing genomic sequence classifications", arXiv preprint arXiv:1605.01133 (2016).

Leaman, R., C.-H. Wei and Z. Lu, "tmchem: a high performance approach for chemical named entity recognition and normalization", Journal of cheminformatics **7**, S1, S3 (2015).

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans", Nature **536**, 7616, 285–291 (2016).

Li, G., D. Jima, F. A. Wright and A. B. Nobel, "Ht-eqtl: integrative expression quantitative trait loci analysis in a large number of human tissues", BMC bioinformatics **19**, 1, 1–11 (2018).

Li, M., Y. Guo, Y.-M. Feng and N. Zhang, "Identification of triple-negative breast cancer genes and a novel high-risk breast cancer prediction model development based on ppi data and support vector machines", Frontiers in Genetics **10**, 180 (2019).

Liao, S., Y. Gao, A. Oto and D. Shen, "Representation learning: a unified deep learning framework for automatic prostate mr segmentation", in "International Conference on Medical image computing and computer-assisted intervention", pp. 254–261 (Springer, 2013).

Lin, L., K. Wang, W. Zuo, M. Wang, J. Luo and L. Zhang, "A deep structured model with radius–margin bound for 3d human activity recognition", International Journal of Computer Vision **118**, 2, 256–273 (2016).

Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken and C. I. Sánchez, "A survey on deep learning in medical image analysis", Medical image analysis **42**, 60–88 (2017).

Liu, L., M. D. Sanderford, R. Patel, P. Chandrashekar, G. Gibson and S. Kumar, "Biological relevance of computationally predicted pathogenicity of noncoding variants", Nature communications **10**, 1, 1–11 (2019).

Liu, X., D. L. Brutlag and J. S. Liu, "Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes", in "Biocomputing 2001", pp. 127–138 (World Scientific, 2000).

Long, J., E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 3431–3440 (2015).

Luo, G., X. Huang, C.-Y. Lin and Z. Nie, "Joint entity recognition and disambiguation", in "Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing", pp. 879–888 (2015).

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery *et al.*, "A systematic survey of loss-of-function variants in human protein-coding genes", Science **335**, 6070, 823–828 (2012).

Mahendran, A. and A. Vedaldi, "Understanding deep image representations by inverting them", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 5188–5196 (2015).

Martins, A. and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification", in "International Conference on Machine Learning", pp. 1614–1623 (2016).

Maston, G. A., S. K. Evans and M. R. Green, "Transcriptional regulatory elements in the human genome", Annu. Rev. Genomics Hum. Genet. **7**, 29–59 (2006).

Maurano, M. T., E. Haugen, R. Sandstrom, J. Vierstra, A. Shafer, R. Kaul and J. A. Stamatoyannopoulos, "Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo", Nature genetics **47**, 12, 1393 (2015).

Narlikar, L. and I. Ovcharenko, "Identifying regulatory elements in eukaryotic genomes", Briefings in functional genomics and proteomics **8**, 4, 215–230 (2009).

Ng, P. C. and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function", Nucleic acids research **31**, 13, 3812–3814 (2003).

Nicolae, D. L., E. Gamazon, W. Zhang, S. Duan, M. E. Dolan and N. J. Cox, "Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas", PLoS Genet **6**, 4, e1000888 (2010).

Nie, D., Y. Gao, L. Wang and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation", in "International Conference on Medical Image Computing and Computer-Assisted Intervention", pp. 370–378 (Springer, 2018).

Ong, C.-T. and V. G. Corces, "Enhancer function: new insights into the regulation of tissue-specific gene expression", Nature Reviews Genetics **12**, 4, 283–293 (2011).

Oubounyt, M., Z. Louadi, H. Tayara and K. T. Chong, "Deepromoter: Robust promoter predictor using deep learning", Frontiers in genetics **10**, 286 (2019).

Ouyang, Z., Q. Zhou and W. H. Wong, "Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells", Proceedings of the National Academy of Sciences **106**, 51, 21521–21526 (2009).

Park, Y. and M. Kellis, "Deep learning for regulatory genomics", Nature biotechnology **33**, 8, 825 (2015).

Peng, Y., K. Yan, V. Sandfort, R. M. Summers and Z. Lu, "A self-attention based deep learning method for lesion attribute detection from ct reports", in "2019 IEEE International Conference on Healthcare Informatics (ICHI)", pp. 1–5 (IEEE, 2019).

Pennacchio, L. A., G. G. Loots, M. A. Nobrega and I. Ovcharenko, "Predicting tissue-specific enhancers in the human genome", Genome research **17**, 2, 201–211 (2007).

Périer, R. C., V. Praz, T. Junier, C. Bonnard and P. Bucher, "The eukaryotic promoter database (epd)", Nucleic acids research **28**, 1, 302–303 (2000).

Polimanti, R. and J. Gelernter, "Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder", PLoS genetics **13**, 2, e1006618 (2017).

Prabhakar, G., B. Kailath, S. Natarajan and R. Kumar, "Obstacle detection and classification using deep learning for tracking in high-speed autonomous driving", in "2017 IEEE Region 10 Symposium (TENSYMP)", pp. 1–6 (IEEE, 2017).

Prastawa, M., J. H. Gilmore, W. Lin and G. Gerig, "Automatic segmentation of mr images of the developing newborn brain", Medical image analysis **9**, 5, 457–466 (2005).

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly *et al.*, "Plink: a tool set for whole-genome association and population-based linkage analyses", The American journal of human genetics **81**, 3, 559–575 (2007).

Qin, Q. and J. Feng, "Imputation for transcription factor binding predictions based on deep learning", PLoS computational biology **13**, 2, e1005403 (2017).

Raj, T., M. Kuchroo, J. M. Replogle, S. Raychaudhuri, B. E. Stranger and P. L. De Jager, "Common risk alleles for inflammatory diseases are targets of recent positive selection", The American Journal of Human Genetics **92**, 4, 517–529 (2013).

Redmon, J., S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 779–788 (2016).

Reimers, N. and I. Gurevych, "Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging", arXiv preprint arXiv:1707.09861 (2017).

Ren, S., K. He, R. Girshick and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", in "Advances in neural information processing systems", pp. 91–99 (2015).

Riethoven, J.-J. M., "Regulatory regions in dna: promoters, enhancers, silencers, and insulators", in "Computational biology of transcription factor binding", pp. 33–42 (Springer, 2010).

Ritchie, G. R., I. Dunham, E. Zeggini and P. Flicek, "Functional annotation of noncoding sequence variants", Nature methods **11**, 3, 294–296 (2014).

Roth, F. P., J. D. Hughes, P. W. Estep and G. M. Church, "Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation", Nature biotechnology **16**, 10, 939–945 (1998).

Sandelin, A., W. Alkema, P. Engström, W. W. Wasserman and B. Lenhard, "Jaspar: an open-access database for eukaryotic transcription factor binding profiles", Nucleic acids research **32**, suppl_1, D91–D94 (2004).

Sargent, D., Y. Yang, N. Šurbanovski, L. Bianco, M. Buti, R. Velasco, L. Giongo and T. Davis, "Haplosnp affinities and linkage map positions illuminate subgenome composition in the octoploid, cultivated strawberry (fragaria× ananassa)", Plant Science **242**, 140–150 (2016).

Schaid, D. J., W. Chen and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping", Nature Reviews Genetics **19**, 8, 491–504 (2018).

Schmidt, F., N. Gasparoni, G. Gasparoni, K. Gianmoena, C. Cadenas, J. K. Polansky, P. Ebert, K. Nordström, M. Barann, A. Sinha *et al.*, "Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction", Nucleic acids research **45**, 1, 54–66 (2017).

Segal, E., Y. Barash, I. Simon, N. Friedman and D. Koller, "From promoter sequence to expression: a probabilistic framework", in "Proceedings of the sixth annual international conference on Computational biology", pp. 263–272 (2002).

Segal, E., T. Raveh-Sadka, M. Schroeder, U. Unnerstall and U. Gaul, "Predicting expression patterns from regulatory sequence in drosophila segmentation", Nature **451**, 7178, 535–540 (2008).

Setio, A. A. A., F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez and B. van Ginneken, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks", IEEE transactions on medical imaging **35**, 5, 1160–1169 (2016).

Shen, D., G. Wu and H.-I. Suk, "Deep learning in medical image analysis", Annual review of biomedical engineering **19**, 221–248 (2017).

Simonyan, K., A. Vedaldi and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps", arXiv preprint arXiv:1312.6034 (2013).

Singh, R., J. Lanchantin, G. Robins and Y. Qi, "Deepchrome: deep-learning for predicting gene expression from histone modifications", Bioinformatics **32**, 17, i639–i648 (2016).

Singh, S., Y. Yang, B. Poczos and J. Ma, "Predicting enhancer-promoter interaction from genomic sequence with deep neural networks", Quantitative Biology **7**, 2, 122–137 (2019).

Sinha, S. and M. Tompa, "A statistical method for finding transcription factor binding sites.", in "ISMB", vol. 8, pp. 344–354 (2000).

Sokolov, A. and A. Ben-Hur, "Hierarchical classification of gene ontology terms using the gostruct method", Journal of bioinformatics and computational biology **8**, 02, 357–376 (2010).

Spielman, R. S., L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens and V. G. Cheung, "Common genetic variants account for differences in gene expression among ethnic groups", Nature genetics **39**, 2, 226–231 (2007).

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting", The journal of machine learning research **15**, 1, 1929–1958 (2014).

Stegle, O., L. Parts, M. Piipari, J. Winn and R. Durbin, "Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses", Nature protocols **7**, 3, 500 (2012).

Strunz, T., F. Grassmann, J. Gayán, S. Nahkuri, D. Souza-Costa, C. Maugeais, S. Fauser, E. Nogoceke and B. H. Weber, "A mega-analysis of expression quantitative trait loci (eqtl) provides insight into the regulatory architecture of gene expression variation in liver", Scientific reports **8**, 1, 1–11 (2018).

Sun, Y., D. Liang, X. Wang and X. Tang, "Deepid3: Face recognition with very deep neural networks", arXiv preprint arXiv:1502.00873 (2015).

Sun, Y. V., Z. Cai, K. Desai, R. Lawrance, R. Leff, A. Jawaid, S. L. Kardia and H. Yang, "Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests", in "BMC proceedings", vol. 1, p. S62 (Springer, 2007).

Tam, V., N. Patel, M. Turcotte, Y. Bossé, G. Paré and D. Meyre, "Benefits and limitations of genome-wide association studies", Nature Reviews Genetics **20**, 8, 467–484 (2019).

Tang, M., P. Gandhi, M. A. Kabir, C. Zou, J. Blakey and X. Luo, "Progress notes classification and keyword extraction using attention-based deep learning models with bert", arXiv preprint arXiv:1910.05786 (2019).

Templeton, A. R., T. Maxwell, D. Posada, J. H. Stengård, E. Boerwinkle and C. F. Sing, "Tree scanning: a method for using haplotype trees in phenotype/genotype association studies", Genetics **169**, 1, 441–453 (2005).

Tomita, N., B. Abdollahi, J. Wei, B. Ren, A. Suriawinata and S. Hassanpour, "Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides", JAMA network open **2**, 11, e1914645–e1914645 (2019).

Toshev, A. and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 1653–1660 (2014).

Trynka, G., C. Sandor, B. Han, H. Xu, B. E. Stranger, X. S. Liu and S. Raychaudhuri, "Chromatin marks identify critical cell types for fine mapping complex trait variants", Nature genetics **45**, 2, 124–130 (2013).

Uçar, A., Y. Demir and C. Güzeliş, "Object recognition and detection with deep learning for autonomous driving applications", Simulation **93**, 9, 759–769 (2017).

Ulirsch, J. C., C. A. Lareau, E. L. Bao, L. S. Ludwig, M. H. Guo, C. Benner, A. T. Satpathy, V. K. Kartha, R. M. Salem, J. N. Hirschhorn *et al.*, "Interrogation of human hematopoiesis at single-cell and single-variant resolution", Nature genetics **51**, 4, 683–693 (2019).

Umarov, R. K. and V. V. Solovyev, "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks", PloS one **12**, 2, e0171410 (2017).

van Arensbergen, J., L. Pagie, V. D. FitzPatrick, M. de Haas, M. P. Baltissen, F. Comoglio, R. H. van der Weide, H. Teunissen, U. Võsa, L. Franke *et al.*, "High-throughput identification of human snps affecting regulatory element activity", Nature genetics **51**, 7, 1160 (2019).

Van de Bunt, M., A. Cortes, M. A. Brown, A. P. Morris, M. I. McCarthy, I. Consortium *et al.*, "Evaluating the performance of fine-mapping strategies at common variant gwas loci", PLoS Genet **11**, 9, e1005535 (2015).

Villar, D., C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska *et al.*, "Enhancer evolution across 20 mammalian species", Cell **160**, 3, 554–566 (2015).

Visscher, P. M., "Sizing up human height variation", Nature genetics **40**, 5, 489–490 (2008).

Wan, X., C. Yang, Q. Yang, H. Xue, X. Fan, N. L. Tang and W. Yu, "Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies", The American Journal of Human Genetics **87**, 3, 325–340 (2010).

Wang, M. and Q. Long, "Addressing issues associated with evaluating prediction models for survival endpoints based on the concordance statistic", Biometrics **72**, 3, 897–906 (2016).

Wei, C.-H., B. R. Harris, H.-Y. Kao and Z. Lu, "tmvar: a text mining approach for extracting sequence variants in biomedical literature", Bioinformatics **29**, 11, 1433–1439 (2013).

Wen, X., Y. Lee, F. Luca and R. Pique-Regi, "Efficient integrative multi-snp association analysis via deterministic approximation of posteriors", The American Journal of Human Genetics **98**, 6, 1114–1129 (2016).

West, A. G., M. Gaszner and G. Felsenfeld, "Insulators: many functions, many mechanisms", Genes & development **16**, 3, 271–288 (2002).

Wittkopp, P. J. and G. Kalay, "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence", Nature Reviews Genetics **13**, 1, 59–69 (2012).

Wu, G., M. Kim, Q. Wang, Y. Gao, S. Liao and D. Shen, "Unsupervised deep feature learning for deformable registration of mr brain images", in "International Conference on Medical Image Computing and Computer-Assisted Intervention", pp. 649–656 (Springer, 2013).

Xie, X., J. Lu, E. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander and M. Kellis, "Systematic discovery of regulatory motifs in human promoters and 3 utrs by comparison of several mammals", Nature **434**, 7031, 338–345 (2005).

Xiong, L., R. Kang, R. Ding, W. Kang, Y. Zhang, W. Liu, Q. Huang, J. Meng and Z. Guo, "Genome-wide identification and characterization of enhancers across 10 human tissues", International journal of biological sciences **14**, 10, 1321 (2018).

Xu, B., X. Shi, Z. Zhao and W. Zheng, "Leveraging biomedical resources in bi-lstm for drug-drug interaction extraction", IEEE Access **6**, 33432–33439 (2018).

Xuan, P., Y. Cao, T. Zhang, R. Kong and Z. Zhang, "Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncrna genes", Frontiers in genetics **10**, 416 (2019).

Yang, J., L. G. Fritsche, X. Zhou, G. Abecasis, I. A.-R. M. D. G. Consortium *et al.*, "A scalable bayesian method for integrating functional information in genome-wide association studies", The American Journal of Human Genetics **101**, 3, 404–416 (2017).

Yang, J., S. H. Lee, M. E. Goddard and P. M. Visscher, "Genome-wide complex trait analysis (gcta): methods, data analyses, and interpretations", in "Genome-wide association studies and genomic prediction", pp. 215–236 (Springer, 2013).

Yuan, L., J. Liu and J. Ye, "Efficient methods for overlapping group lasso", in "Advances in neural information processing systems", pp. 352–360 (2011).

Yuan, Y., L. Guo, L. Shen and J. S. Liu, "Predicting gene expression from sequence: a reexamination", PLoS Comput Biol **3**, 11, e243 (2007).

Zaykin, D. V. and L. A. Zhivotovsky, "Ranks of genuine associations in whole-genome scans", Genetics **171**, 2, 813–823 (2005).

Zeiler, M. D. and R. Fergus, "Visualizing and understanding convolutional networks", in "European conference on computer vision", pp. 818–833 (Springer, 2014).

Zeng, B., L. R. Lloyd-Jones, A. Holloway, U. M. Marigorta, A. Metspalu, G. W. Montgomery, T. Esko, K. L. Brigham, A. A. Quyyumi, Y. Idaghdour *et al.*, "Constraints on eqtl fine mapping in the presence of multisite local regulation of gene expression", G3: Genes, Genomes, Genetics **7**, 8, 2533–2544 (2017).

Zeng, B., L. R. Lloyd-Jones, G. W. Montgomery, A. Metspalu, T. Esko, L. Franke, U. Vosa, A. Claringbould, K. L. Brigham, A. A. Quyyumi *et al.*, "Comprehensive multiple eqtl detection and its application to gwas interpretation", Genetics **212**, 3, 905–918 (2019).

Zeng, J., R. De Vlaming, Y. Wu, M. R. Robinson, L. R. Lloyd-Jones, L. Yengo, C. X. Yap, A. Xue, J. Sidorenko, A. F. McRae *et al.*, "Signatures of negative selection in the genetic architecture of human complex traits", Nature genetics **50**, 5, 746–753 (2018).

Zhang, X., D. D. Wu, X. Zhou and X. Hu, "A language modeling text mining approach to the annotation of protein community", in "Sixth IEEE Symposium on BioInformatics and BioEngineering (BIBE'06)", pp. 12–19 (IEEE, 2006).

Zhang, Y. and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies", Nature genetics **39**, 9, 1167–1173 (2007).

Zhao, H. and A. Dean, "An insulator blocks spreading of histone acetylation and interferes with rna polymerase ii transfer between an enhancer and gene", Nucleic acids research **32**, 16, 4903–4919 (2004).

Zhou, J., 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", Nature genetics **50**, 8, 1171–1179 (2018).

Zhou, J. and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model", Nature methods **12**, 10, 931–934 (2015).

Zintgraf, L. M., T. S. Cohen, T. Adel and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis", arXiv preprint arXiv:1702.04595 (2017).