Whole genome regulatory variant evaluation for transcription factor binding

by

Haoyang Zeng

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
May 15, 2015

Signature redacted

Certified by .....
Professor David K. Gifford
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by .....
Professor Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

Whole genome regulatory variant evaluation for transcription factor binding

by

Haoyang Zeng

Submitted to the Department of Electrical Engineering and Computer Science
on May 15, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

With the advent of high-throughput sequencing technology, Genome Wide Association Studies (GWAS) have identified thousands of genetic variants that are associated with disease and complex traits. Many of these variants reside in the non-coding region of the genome, and affect gene expression and downstream cellular phenotype by disrupting the regulatory machinery of the cell. For example these variants can alter the binding of the transcription factors (TF). In this thesis we present Whole-genome regulAtory Variant Evaluation (WAVE), a computational method that models the TF binding ChIP-seq signal solely from DNA sequence and predicts genetic a variant's effect on TF binding. Applying WAVE to two important transcription factors, NF-κB and CTCF, we show that WAVE accurately predicts ChIP-seq signal on held-out chromosome. WAVE discovers the DNA motif of the target TF as well as the binding co-factors, displaying substantially greater expressiveness in modeling TF binding than conventional motif-based approaches. Furthermore, with AUC larger than 0.7 in the most stringent control scenario, WAVE outperformed existing motif-based approaches in predicting genetic variants associated with allele-specific binding.

Thesis Supervisor: Professor David K. Gifford
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I find it a great privilege to work with the brilliant faculty and students at MIT, which provides me with not only fruitful results in research but more importantly the faith to aim at some of the most difficult problems in genomics, the right methodology to dissect and approach problems, and the consistent self-interrogation on the research quality. I am deeply thankful for the research guidance, availability and mental support from my adviser Prof. David Gifford who will always be there for his students. I thank Tatsu Hashimoto for his patient mentoring during my rotation in the lab and great resourcefulness in the course of this work. I thank Matthew Edwards for the insightful discussions and patient help with working on the lab cluster. I appreciate the help from Yuchun Guo in performing analysis using GEM. I thank other members of the Gifford lab for constructive suggestions and useful feedback on the thesis. Finally, I am deeply thankful to my family, friends and particularly Menghan for consistent support throughout my research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Next-generation sequencing (NGS) technology has dramatically decreased the cost and time for whole-genome sequencing, enabling the possibility of large-scale genotying of human individuals across various population cohorts. Exploiting the power of the technology, Genome-Wide Association Studies (GWAS) provide a systematic method to examine the possible associations between common genetic variants and complex traits such as major diseases [1–4]. Thousands of genetic variants have been identified by GWAS studies to be significantly associated with important diseases from schizophrenia to Alzheimer's disease. However, the mechanism by which these variants disrupt cellular function and contribute to downstream phenotypes remains largely unknown.

Variants in protein coding sequences including missense and nonsense mutations are simple to characterize, as they have direct effects on the composition of the protein transcribed from the coding sequence. However, the majority of GWAS detected variations reside in non-coding regions with potential regulatory roles in the cell machinery [4, 5]. With the lack of functional annotation in these non-coding regions, the influences of such variation on gene expression and other cellular functions are not well understood.

The gene expression in the cell is regulated through complicated machinery that

involves the interaction of many protein complexes, and transcription factors (TF) play an important regulatory role in these complexes. Previous work has observed that non-coding DNA changes in the recognition sequences of transcription factors can affect gene expression and cellular phenotypes [6]. Thus predicting the effect of genomic variants on TF binding is an essential for interpreting the role of non-coding variants in pathogenesis.

1.2 Experimental methods for regulatory variant discovery

Several generations of experimental approaches have been designed to determine the transcription binding site *in vivo*, among which chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) is a widely used method . In the ChIP-seq protocol, the target transcription factor is cross-linked *in vivo* to DNA which is then sheared to fragments using sonication. These DNA fragments are co-precipitated with transcription factor-specific antibodies, unlinked and purified for high-throughput sequencing. Mapping the sequencing reads to the genome result in a genome-wide profile of the binding of the target transcription factor.

After obtaining *in vivo* binding data from ChIP-seq experiments, two categories of approaches are usually used to determine the effect of genetic variants on transcription factor binding.

Association Test

An association test exploits the statistical power of the large number of genotype and ChIP-seq data from different individuals. Referring to the haplotype without the variant as reference allele and the haplotype with the variant as the alternate allele, the association test is built on the assumption that the total binding strength of the TF around a variant site should be associated with the number of alternate alleles (0,1 or 2 for human as we are diploid creature) if that variant has an causal impact

on the binding signal.

Allele-Specific Analysis

Allele-specific analysis is built on the assumption that heterozygous variants will exhibit an "allelic imbalance" where the reads will come from one allele more often than the other depending on whether the variant has a positive or negative effect on the TF binding.

Although experimental methods that utilize high-throughput sequencing technology such as ChIP-seq provide direct evidence for the effect of genetic variants on transcription factor binding, it is limited in several aspects. First, both of the two types of analysis requires ChIP-seq experiments with a reliable antibody to the target transcription factor, which is not available for some of the factors. Second, an association-test analysis needs genotype and ChIP-seq data from a large number of individuals, while allele-specific analysis requires very deep sequencing to have sufficient read depth for reliable identification of allelic imbalance. Both of these requirements are costly and time-consuming. Thirdly, many variants are excluded from the analysis either because they are not common enough to appear in the limited number of samples in the association-test analysis, or because they are not close enough to the peak of the binding (therefore without enough read depth) in the allele-specific analysis. Both of these factors limit the range of variants that we are able to evaluate. Therefore, developing models that could computationally predict the effect of variant becomes crucial for the genome-wide discovery and assessment of regulatory variant .

1.3 Computational methods for regulatory variant discovery

Because of the limitation of experimental methods discussed above, a series of computational approaches to evaluate the effect of genetic variants have been proposed [7–13]. However, they are all based on the classical model of transcription factor binding

using sequence motif, which severely limits their efficiency and scope in regulatory variant discovery.

1.3.1 Modeling transcription factor binding using motif

When binding to DNA, majority of the transcription factors recognize and bind to a factor-specific sequence of 6-20 bp long. The binding preference of each transcription factor can be summarized as a consensus matrix or 'position weight matrix' (PWM) where at some positions (called 'degenerated positions') different nucleotides are allowed. This representation is called a 'binding motif'. Different experimental and computational methods of determining the binding motifs for transcription factors have been developed, and the binding preference found by different studies have been systematically collected in databases such as JASPAR, TRANSFAC and UniProbe.

The simplicity of motif representation of individual transcription factors binding to their preferred binding sequences enables efficient scanning of putative binding site on the entire genome based on how well the sequence matches the PWM or so called "PWM score". However, it has been often criticized for its inability of capturing some of the important information in transcription factor binding, such as the non-canonical motifs, the role of binding co-factors and interaction between nucleotides inside motif.

1.3.2 Prior work on computational models

Existing computational methods are all based on quantifying how differently the reference and alternate allele of a variant match with the binding motif. In terms of the category of the variant, they all focus on single nucleotide polymorphisms (SNPs), the simplest form of genetic variants. RAVEN [7] scores a SNP by the change of the PWM score on the two alleles. is-rSNP [8] scores a SNP by the ratio of p-values for the PWM score on the two alleles. sTRAP [9] scores a SNP by the ratio of p-values of a Fourier transform derived affinity score of the two alleles. rSNP-MAPPER [10] scores a SNP by the difference of the two allele matching with binding motif represented in

a Profile Hidden Markov Model. HaplogReg2 [11] scores a SNP as altering the TF motif when at least one of the alleles passes a pre-defined p-value threshold of the PWM matching score.

1.4 Thesis Overview

The rest of the thesis are organized in the following manners. Chapter 2 describes a novel computational approach to predict the effect of genetic variants on TF binding. Chapter 3 investigates the properties of this model and compares it against existing computational models under the same topic. Chapter 4 summarizes the thesis and discusses possible future extensions.

Chapter 2

Methods

Previous work [14] successfully predicted DNase-seq signal from DNA sequence alone, using a computational framework (referred as k-mer model in this thesis) that maps from sequence to high-throughput sequencing data such as DNase-seq read counts.

Several properties of the k-mer model make it a good choice for modeling the binding of transcription factor binding and predicting the effect of genetic variants on the binding signal:

- **Spatial effect** The ability of the k-mer model to capture the spatial effect of DNA sequence on read count data makes it possible to model the behavior of binding co-factors and other auxiliary sequences in TF binding. This capability is beyond the power of existing motif-based methods.

- **Great expressiveness** The k-mer model doesn't assume the existence of a canonical motif, but instead concurrently models the effect of all the possible k-mers with parameter regularization. Therefore, it has greater ability in modeling non-canonical motif and other more subtle sequence features that control TF binding.

- **Parameter Regularization** The k-mer model uses L1-regularization to enforce the sparsity of its parameters. This allows us to identify the key set of k-mers that are crucial for TF binding.

- **Applicability to other high-throughput data type** Although the k-mer model was originally proposed [14] as a mean to model DNase-seq data, the model is not limited to model any specific data type. This enables the possibility of adopting to model other high-throughput sequencing data such as ChIP-seq.

Given these good properties, we adopted the k-mer model to predict ChIP-seq signal. We score a genetic variant by the change of the predicted signal due to the variant. In this chapter, we first describe the methodology behind k-mer model and then introduce the full pipeline of WAVE model. Lastly we describe some of the method used in the analysis of the WAVE model.

2.1 K-mer model

A k-mer model considers the genome as a long regulatory sequence that contains k-mer as "code words" that induce invariant spatial effects on proximal transcription factor binding. Following this assumption, it models the read counts at a given base on the genome as the log-linear combination of spatial effect of a set of learned k-mers whose effect range covers that base.

The effect profile of a k-mer is defined as a real-valued vector of length $2M$ that corresponds to a spatial window of $[-M, M-1]$ relative to the start position of the k-mer. Specifically, the j-th entry of the profile for a k-mer is the expected log-change in read counts at the j-th base relative to the start of the k-mer. For notational convenience we will use i for genomic coordinate, k for k-mer length, and j for coordinate offset from the start of a k-mer. We assume that the genome consists of one large chromosome with coordinate 0 to N . In practice we will construct this by concatenating chromosomes with the telomeres acting as a spacer. We represent the effect vector of all k-mer of length k as a parameter matrix θ^k of size $4^k \times 2M$. For any particular k-mer of length k starting at base i on the reference genome, we define g_i^k as its row index in θ^k . So $\theta^k_{g_i^k,j}$ would denote the effect of this kmer at offset $j \in [-M, M-1]$. Additionally, a special parameter θ_0 is used to set the average read rate of the genome globally.

Given these definitions, a generative model for high-throughput reads on the genome was proposed in which observed counts c_i are generated from a Poisson distribution with rate parameter λ_i defined as:

$$\lambda_i = \exp((\sum_k \sum_{j \in [-M-k_{max}+1, M-1]} \theta^k_{(g^k_i, j)}) - \theta_0) \tag{2.1}$$

The problem to solve then is a regularized Poisson regression. Particularly, we would like to maximize the following:

$$\max_\theta \{(\sum_i c_i \log(\lambda_i) - \lambda_i) - \eta \sum \left\|\theta^k\right\|_1\} \tag{2.2}$$

To efficiently optimize this objective function, the model uses minibatch-gradient descent, a variant of stochastic gradient descend method where in each iteration the gradient and error is calculated against a 'mini-batch' of all the samples [15]. The detail of implementation can be found in the supplementary text of [14].

2.2 Overview of WAVE model

The WAVE pipeline of variant scoring consists of the following three steps (Figure 2-1):

- Learn the spatial effect of k-mers on TF binding from ChIP-seq datasets.
- Predict the TF binding signal around the reference and alternate allele of each SNP of interest
- Score a SNP by the sum of squared per-base change of binding signal between alleles

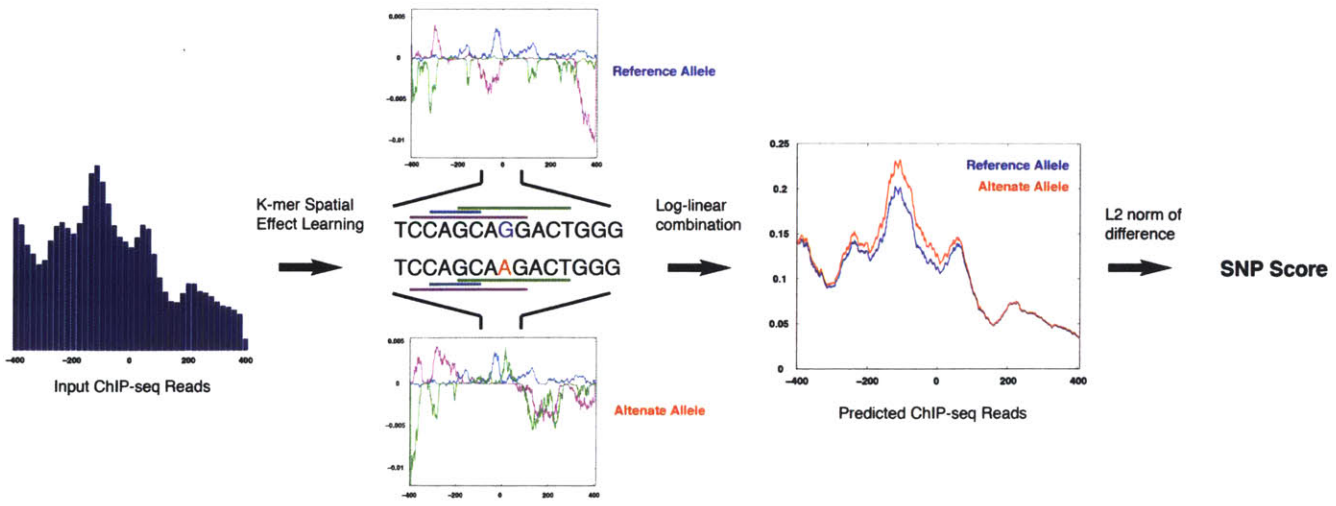


Figure 2-1: The schematic of WAVE pipeline

2.3 K-mer spatial profile learning

Using k-mer model, WAVE first learns the spatial effect of all k-mers (k=1 to 8) over a spatial window of ±400 base pairs (bp) de novo from ChIP-seq data. Here we consider k-mers with k from 1 to 8 ($k_{max} = 8$) as this is the maximum length learnable with a typical ChIP-seq dataset. As ChIP-seq signals are relatively sparse and spikey, we chose an effect range of ±400 bp for each k-mer ($M = 400$) instead of ±1000 bp ($M = 1000$) as was used in the original k-mer model used for modeling DNase-seq signal [14]

2.4 Predicting the ChIP-seq signal for reference and alternate allele

For each genetic variant, we refer to the original allele without the variant as reference allele, and the allele with the variant as alternate allele. The generative nature of the k-mer model enables the possibility of quantitatively characterizing the binding change between the reference and the alternate allele. Specifically, if the variant is located at position i on the genome, then k-mers reside in the range of $[i - k_{max} + 1, i]$ will be affected and therefore the binding signal in the range of $[i - k_{max} + 1 - M, i +$

$M-1]$ will be altered by the variant.

Given the effect profile θ^k of all the k-mers learned from k-mer model trained in step 1, we then predict the expected ChIP-seq count λ_i at each position i across the reference genome by combining the effect of proximal k-mers into the log-linear model using equation 2.1. In a similar manner, we predict the read counts of the alternate allele $\lambda_i^{'}$ after replacing the k-mers that are affected by the variant. If we assume a Single Nucleotide Polymorphism (SNP), at most $\frac{4}{3} \times (4^{k_{max}} - 1)$ k-mers will change.

2.5 Variant scoring

In step 3, we score a SNP at locus i on the genome by the sum of squared per-base change of predicted binding signal at all bases within the effect range of any k-mers affected by the variant:

$$s_i = \sum_{j \in [-M-k_{max}+1, M-1]} (\lambda_{i+j}^{'} - \lambda_{i+j})^2 \tag{2.3}$$

2.6 Analysis of the performance of WAVE

In this section, we describe the detail in how we analyze the performance of WAVE and compare it with existing methods.

2.6.1 Discriminating SNPs known for altering TF binding from negative control

As was introduced in detail in Chapter 3, we performed a series of analysis to benchmark the performance of WAVE and other existing motif-based methods in distinguishing positive SNPs reported to alter TF binding from negative controls set. In all the analysis of this kind, we used WAVE and other competing methods to numerically score all the variants in the positive and negative set. For each method, we varied the score cutoff to produce a receiver operating characteristic (ROC) curve reflecting how well the method prioritizes the positive set over the negative control.

Then we quantified the performance by calculating the area under the curve (AUC) of the ROC, which would be 0.5 for random chance and 1 for a perfect classifier.

2.6.2 Comparing with existing motif-based methods

sTRAP

We used the R version of sTRAP downloaded from the website for scalability. We used motif data from the JASPAR (included in sTRAP R package) and TRANSFAC (2013.1) databases, including MA0105.1, MA0107.1, MA0061.1, M00054, M00194, M00052, M00051, M03557, M00208, M03563 for NF-κB and MA0139.1, M01200, M01259 for CTCF. For binary classification, we chose an absolute log ratio cutoff of 0.21 and min-pvalue cutoff of 0.1 as was suggested by the sTRAP paper. To plot the sTRAP ROC curve we ranked the SNPs by their absolute log ratio.

rSNP-MAPPER

We scored SNPs with rSNP-MAPPER using the models associated with target TF in rSNP-MAPPER model library, including MA0105, MA0107, MA0061, M00774, M00054, M00052, T00595, T00594, T00606, T00593, T00592, T00591, T00588, T00587, T00590, M00051 for NF-κB and T00284 for CTCF. For binary classification, we used a score cutoff of 0 and score change cutoff of 2, as suggested in the rSNP-MAPPER paper. To plot the rSNP-MAPPER ROC curve we ranked the SNPs by their score change.

HaploReg2

We used HaploReg2's default parameters. As HaploReg2 is not able to give a numeric score for each SNPs, we performed binary classification of each SNP by looking for "NF-kappaB" or "CTCF" in the Motif column of the result for SNP sets of NF-κB and CTCF respectively.

2.6.3 Collapsing WAVE Profiles into PWM

We interpret the active k-mers captured by WAVE with a post-processing framework that aggregates similar k-mers into position weight matrices after filtering on effect size. The frame work consists of following steps:

- We filter k-mers based on the sum of effect to eliminate inactive k-mers.
- We calculate the pair-wise Levenstein distance of the remaining k-mers.
- We perform UPGMA hierarchical clustering over the candidate k-mers until the minimal distance among clusters is larger than 2.
- For each cluster, we define its key k-mer as the one with the largest aggregate effect. We obtain the position weight matrix for this cluster by aligning all k-mers in the cluster against the key k-mer.

Chapter 3

Results

3.1 WAVE learns a vocabulary of k-mers that regulate factor binding

We first tested if WAVE could predict held-out ChIP-seq data. We trained a WAVE model on NF-κB ChIP-seq data from chromosomes 1-13 of 10 LCL ENCODE individuals and compared the predicted ChIP-seq signal from WAVE to ChIP-seq reads on a held-out chromosome (chromosome 14). The predicted ChIP-seq signals are very similar to actual ChIP-seq reads (Figure 3-1), with Pearson's correlations of 0.64 chromosome-wide and 0.44 restricted to regions within 2kb of a binding event identified by GEM [16]. Negative control WAVE models trained to capture biases such as chromatin state on rabbit IgG ChIP-seq datasets yield Pearson's correlations of only 0.51 chromosome-wide and 0.10 within 2kb of binding events. The correlation values are not as high as those reported for DNase-seq in the original paper [14] which yields a correlation of 0.801 after filtering out low-signal regions. This could be because ChIP-seq data are much sparser and spikier than DNase-seq, and therefore harder to model in terms of magnitude.

As a complementary analysis, we calculated the standardized original and predicted ChIP-seq signal (z-score) for each base on the genome. The z-scores of original ChIP-seq read counts have a much better linear correlation with z-scores of signal

predicted by WAVE than with the signal predicted by the negative control WAVE model (Figure 3-2).

To further investigate how well WAVE captures the peaks of ChIP-seq signal, we compared the binding calls from GEM taking the original and WAVE-predicted ChIP-seq signal respectively as experiment input (referred as original GEM and predicted GEM in this thesis). The rabbit IgG ChIP-seq data was used as control input to GEM in both cases. As each binding event called by GEM is a single-base position on the genome, we defined the peaks as the region within 200 bp around each binding events. We further define coverage as the ratio of peaks called by original GEM that was also called by predicted GEM, and positive predicted value (ppv) as the ratio of peaks called by predicted GEM that was also called by original GEM. We achieved a coverage of 0.17 and ppv of 0.31 genome-wide, showing that WAVE captures around 20% of the real peaks. These results are consistent with the coverage and ppv results in the original paper [14] where k-mer model was used to model DNase-seq data. Many factors could lead to the failure of capturing many of the peaks, such as the non-specificity of the antibody used in the ChIP-seq experiment and the indirect binding of NFKB with the help of other co-factors. The fact that k-mer model caps and smooths the ChIP-seq data before learning and therefore leads to a predicted signal that is less spiky than real ChIP-seq signal might also confound the peak calling using GEM.

Although WAVE fits a model with potentially large parameter space ($\pm$400 bp window for 87380 k-mers when max k $\geq$ 8), it uses sparsifying regularization to avoid over-fitting and to limit the number of active k-mers. For example, in the NF-κB WAVE model, most of the binding signal is predicted by 1% of the 87380 k-mers(Figure 3-3). WAVE is also robust to the choice of window sizes of k-mer's spatial effect,although we found that WAVE model with window size of $\pm$400bp produced the best Pearson's correlation (Table 3.1)

We also found the SNP scores generated by WAVE are consistent across similar training datasets. We trained four separate WAVE models on NF-κB ChIP-seq data from four different individuals GM12878 (CEU), GM12892 (CEU), GM18951 (JPT)

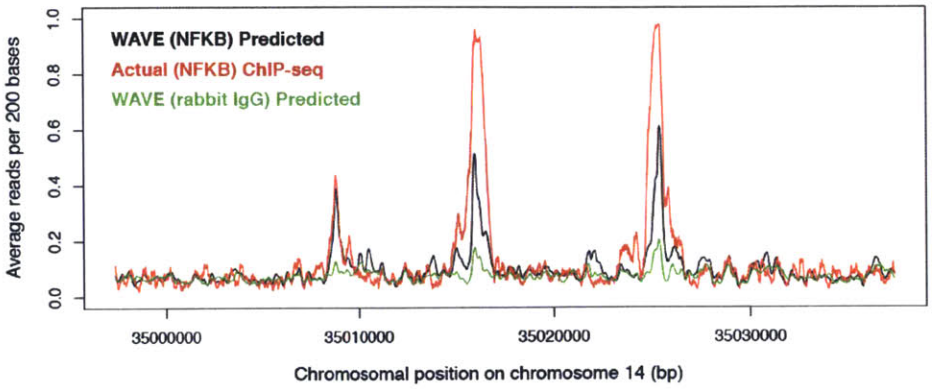


Figure 3-1: Example held-out genomic region on chromosome 14 showing ChIP-seq reads (red), WAVE-predicted reads (black), and reads from a negative control model trained on rabbit IgG ChIP-seq data (green).

TF	Data Source	K-mer Effect Window Size	Pearson's Correlation
CTCF	GM12878	800	0.52
CTCF	10 LCL individuals Combined	800	0.62
NFKB	GM12878	800	0.55
NFKB	10 LCL individuals Combined	800	0.64
NFKB	10 LCL individuals Combined	400	0.63
NFKB	10 LCL individuals Combined	2000	0.59

Table 3.1: Summary of chromosome-wide Pearson's correlation under different experiment setup

and GM19193 (YRI). The SNP scores for the set of common (minor allele frequency $\geq$1%) SNPs from the 1000 Genomes Project (1KG) are consistent across the four different models (Figure 3-4). Moreover, we found the WAVE model trained on the combined ChIP-seq data from 10 LCL individuals had a clear improvement in Pearson's correlation between predicted and actual reads when compared with any model trained from a single individual (Table 3.1)

3.2 WAVE captures primary and auxiliary sequences

We then examined if WAVE correctly learned the strongest expected sequence features from the binding data that correspond to the canonical motifs for NF-κB and CTCF.

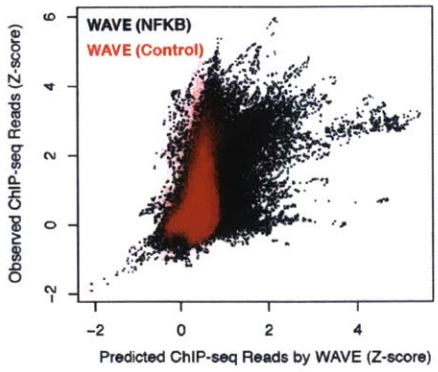


Figure 3-2: Comparison of z-score of WAVE-predicted (x-axis) and observed (y-axis) ChIP-seq reads in binned regions of held-out chromosome 14. Models were trained on combined ChIP-seq data from 10 ENCODE LCL individuals (black) or rabbit IgG ChIP-seq data (red).

Both WAVE models were trained on combined ChIP-seq data from 10 LCL ENCODE individuals and position weight matrices were generated for visualization purposes by hierarchical clustering of the active k-mers in WAVE and matched to known TF motifs in JASPAR and TRANSFAC with STAMP [17]. We found that the top two k-mer clusters for NF-κB were strongly matched to motifs from NF-κB family (Figure 3-5(a)) and the top 6 k-mer clusters for CTCF were all strongly matched to the CTCF motif (Figure 3-5(b) and Figure 3-6).

Many of the other k-mer clusters learned by WAVE correspond to co-factor binding motifs. The top k-mer clusters in the NF-κB WAVE model matched to ETS1, AP1, IRF1 and NRF1 (Figure 3-5(c)), which have been associated with NF-κB regulation [18–21] . To validate the role of these transcription factors in NF-κB binding, we performed co-factor analysis on the same NF-κB data using GEM to search for transcription factors that have spatially binding constraint with NF-κB. This analysis identified AP-1 and IRF1 as the strongest co-factors of NF-κB binding. Thus, WAVE captures the sequence context of factor binding and provides additional descriptive power.

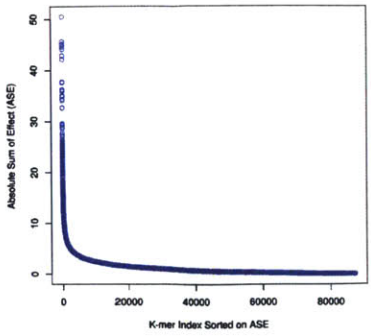
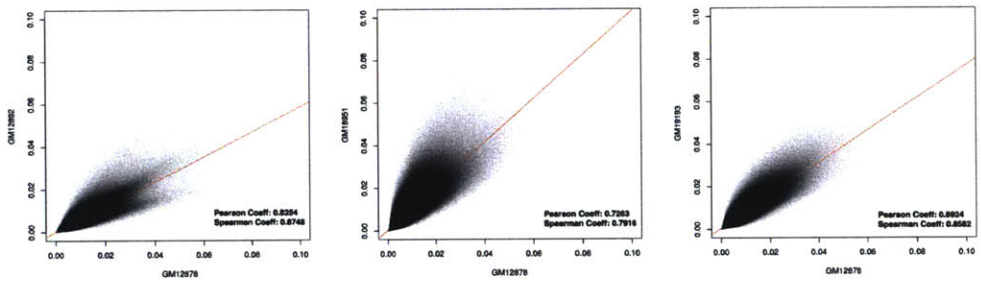


Figure 3-3: Distribution of the summed effect-size of all 87380 k-kmers with length less than or equal to 8 from NF-κB WAVE model

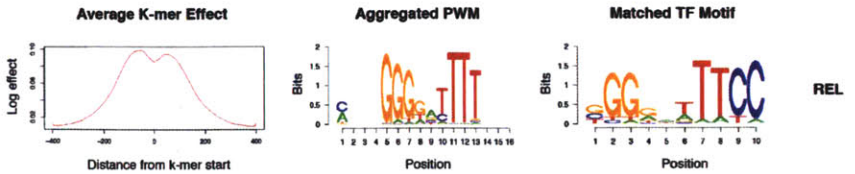


(a) GM12878 vs. GM12892 (b) GM12878 vs. GM18951 (c) GM12878 vs. GM19193

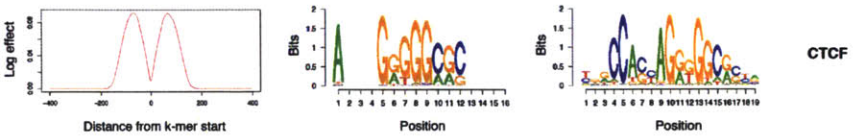
Figure 3-4: WAVE is consistent cross LCL cell lines. Scatter plots of SNP scores assigned by WAVE models trained on ChIP-seq data from different LCL individuals.

3.3 WAVE outperforms motif-based methods in predicting ASB SNP

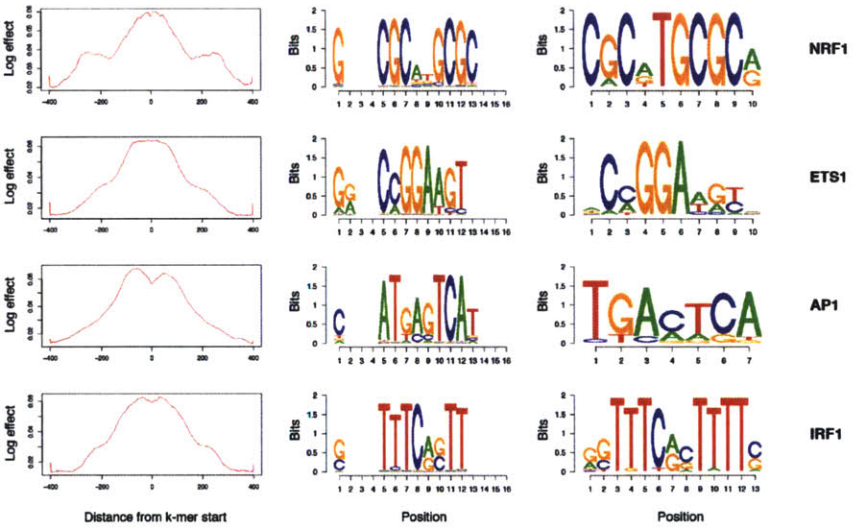
We then compared WAVE's performance against motif-based approaches in discriminating SNPs that are known to alter transcription factor binding. Allele-specific binding (ASB) studies have identified SNPs associated with significantly imbalanced binding events on heterozygous sites, making these SNPs an ideal standard for benchmarking [22, 23]. We used the list of SNPs that are reported to induce allele-specific binding (ASB) of NF-κB and CTCF in GM12878. Our NF-κB positive SNP set consists of 70 ASB SNPs combined from [22] and [23]. Our CTCF positive SNP



(a) Strongest Motif for NFKB



(b) Strongest Motif for CTCF



(c) Top k-mer clusters for NFKB captures motif for binding co-factors

Figure 3-5: Active k-mers detected by WAVE contains the canonical TF motif and associated cofactors

set consists of 1336 ASB SNPs from [22]. After filtering on minor allele frequency ($\geq$0.01), we are left with 54 SNPs for NF-κB and 1247 SNPs for CTCF, from the latter of which we further down-sampled 60 SNPs as our final CTCF positive SNP set to accommodate the limited scoring throughput of motif-based approaches evaluated in this study.

We constructed three sets of negative SNPs that we assume do not exhibit differential factor binding. All of these negative sets are subsets of 1KG common (minor allele frequency $\geq$1%) SNPs. The first negative set is a random selection of 1000

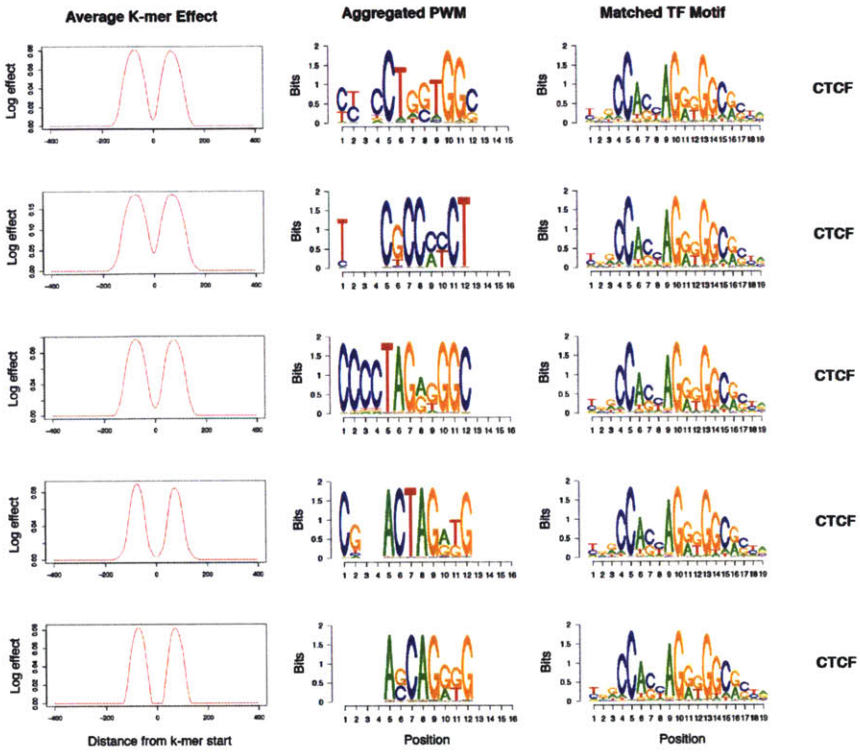
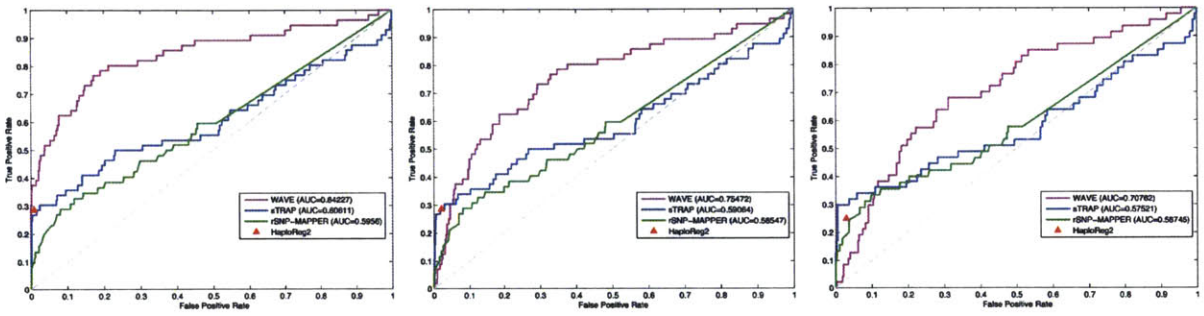


Figure 3-6: The 2^{nd} to the 6^{th} strongest k-mer clusters for CTCF captures different parts of long CTCF motif. The average k-mer effect and aggregated position weight matrix of the 2rd, 3rd, 5th, 6th and 7th strongest k-mer cluster from CTCF WAVE model compared with CTCF motif (MA0139.1).

1KG common SNPs from across the genome in order to sample overall background. To account for the non-uniform distribution of ASB SNPs on the genome, the second negative set was composed of 1KG common SNPs within 1kb from an ASB SNP. We found that 47 out of 56 NF-κB ASB SNPs reside in one of the 15522 NF-κB binding regions (BR) identified by previous work [24]. Thus the third negative set was constructed for NF-κB analysis only to control for the confounding effects arise from proximal binding strength. This final negative set is a subset of second negative set that are located in any BR that contains a positive ASB SNP.

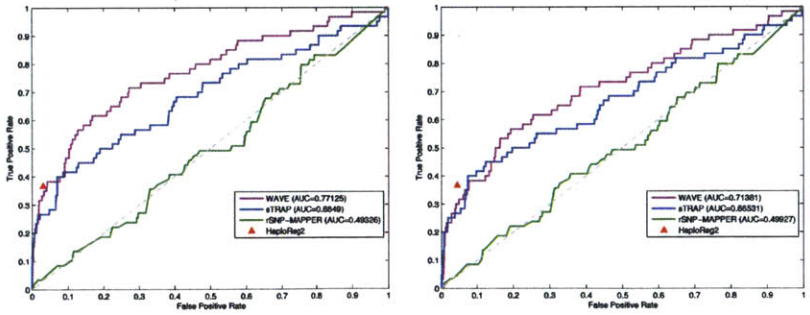
We evaluated the performance of HaploReg2, sTRAP, rSNP-MAPPER, and WAVE in discriminating our positive set from each of our three negative sets. The NF-κB and CTCF WAVE models were trained on combined data from 10 LCL ENCODE individuals. The other models are evaluated as described in Chapter 2

We found that WAVE outperformed all the other tested methods and had an AUC > 0.7 on the third and most stringent negative set (Figure 3-7 and Figure 3-8). Since HaploReg2 does not provide a way to retrieve the actual score for each SNP, its binary classification performance was plotted as a point on the ROC curves.



(a) Control set 1 (b) Control set 2 (c) Control set 3

Figure 3-7: WAVE significantly outperforms motif-based approaches in prioritizing NF-κB ASB-SNPs. ROC curves for discriminating NF-κB ASB-SNPs from each of the three negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance.



(a) Control set 1 (b) Control set 2

Figure 3-8: WAVE significantly outperformed motif-based approaches in prioritizing CTCF ASB-SNPs. ROC curves for discriminating CTCF ASB-SNP from the various negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance.

3.4 WAVE prioritizes SNPs that disrupt motifs

The power of motif-based approaches is constrained to evaluating variants that fall inside a factor's motif. To evaluate the power of WAVE on this narrower task we next tested WAVE's performance on SNPs in our positive set that could be detected by motif based methods.

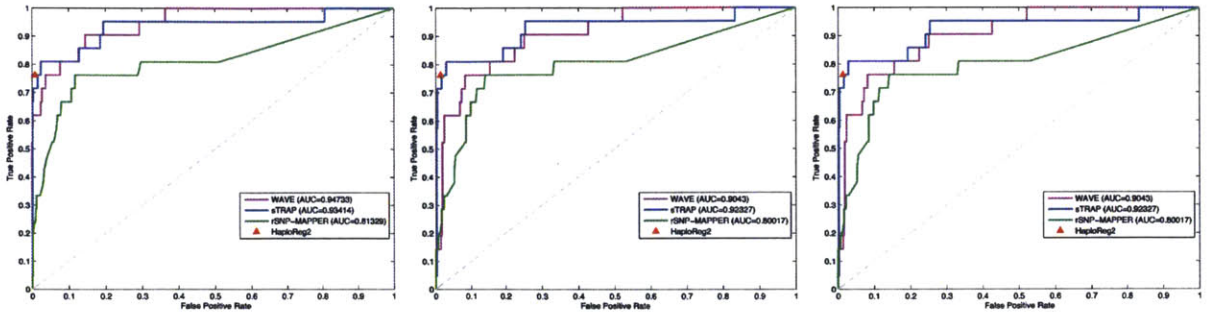
As the first step, we classified our positive sets of NF-κB and CTCF ASB SNPs using sTRAP, rSNP-MAPPER and HaploReg2 with their default parameter and cutoff settings. As motif disruption has been considered the primary mechanism by which variants alter transcription factor binding, we would expect most of the ASB SNPs to be identified by these methods. Surprisingly we found that the best motif-based method (rSNP-MAPPER for NF-κB and HaploReg2 for CTCF) detected only 30% of our positive SNPs. The set of correctly classified positive SNPs by the approaches significantly overlapped, while the efficiency of each method varied across different transcription factors (Table 3.2). A total of 21 out of 56 NF-κB ASB SNPs and 24 out of 60 CTCF ASB SNPs were detected by at least one of the three motif-based methods. We refer these SNPs as motif-disrupting (MD) SNPs. A large fraction (around 60%) of positive SNPs were not detected by any of the traditional motif-based approaches. These results are consistent with our previous results (Figure 3-7) which shows that the prediction power of motif-based approaches dramatically decreases to random after the top 30% positive targets.

Method	NF-κB	CTCF
HaploReg2	16/56	22/60
sTRAP	12/56	10/60
rSNP-Mapper	19/56	2/60
Combine above	21/56	24/60

Table 3.2: Existing motif-based methods only detects 30% ASB SNPs

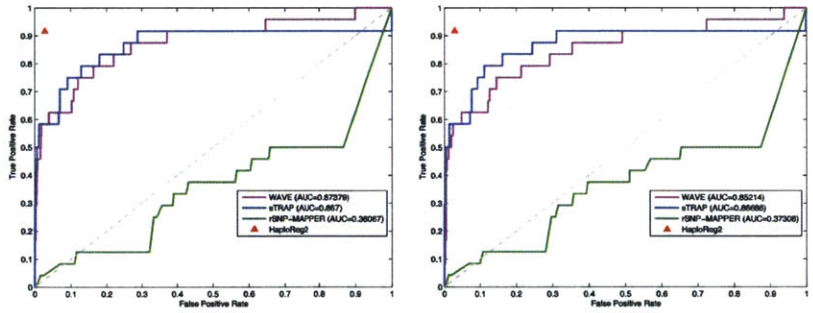
We then compared the performance of WAVE, sTRAP, rSNP-MAPPER and HaploReg2 in discriminating the motif-disrupting SNPs from our three negative sets. For both transcription factors, WAVE achieved performance equal to motif-based ap-

proaches with an AUC >0.85 (Figure 3-9, Figure 3-10) in all control scenarios.



(a) Control set 1 (b) Control set 2 (c) Control set 3

Figure 3-9: WAVE achieves performance equal to motif-based approaches in prioritizing NF-κB SNPs within binding motifs. ROC curves for discriminating NF-κB SNPs identified by any motif-based method (MD SNPs) from the various negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance.



(a) Control set 1 (b) Control set 2

Figure 3-10: WAVE achieved similar performance as motif-based approaches in prioritizing CTCF MD SNPs. ROC curves for discriminating CTCF MD SNPs from the various Negative sets using WAVE, sTRAP, rSNP-MAPPER and HaploReg2. Dashed line indicates random chance.

Chapter 4

Conclusion

We have proposed a computational model WAVE that accurately predicts the effect of genetic variants on transcription factor binding. Unlike existing methods that are based on binding motif discovery, WAVE directly predicts ChIP-seq signal from DNA sequence and scores a variant by the change of predicted binding signal between the reference and alternate allele.

We have found the power of position weight matrices to be insufficient to properly score the effect of variants on factor binding. Motif-based approaches were only able to correctly annotate 30% of ASB SNPs in our test set. In addition, the performance of a motif-based model can vary dramatically across different transcription factors (Table 3.2). We expect that the poor performance of rSNP-MAPPER on CTCF might be the consequence of its use of a single CTCF position weight matrix. Thus motif-based methods are strongly constrained by their underling model.

WAVE significantly outperformed motif-based scoring methods in prioritizing ASB SNPs from negative controls. We find that WAVE's incorporation of a window of sequence context permits it to model the effect of other sequences involved in transcription factor binding including co-factors. These sequences are neglected by conventional motif-based motif modeling.

WAVE outputs a numeric score for each SNP that is easy to interpret as the predicted number of reads changed by the variant. We showed that WAVE scores are robust to the choice of window size, and are consistent across the individuals used for

training. We further demonstrated that by combining ChIP-seq data from multiple individuals of the same cell lines type to increase the size of the training set, we could improve WAVE's goodness of fit to ChIP-seq reads

4.1 Contributions

The primary contributions of this works are:

- The first fully generative model that directly predicts ChIP-seq signal from DNA sequence. Not only does this produce scores of variants that can be straightforwardly interpreted as the predicted change of binding signal, it also allows us to understand how exactly each k-mer affect the binding signal around it.

- Concurrent modeling of all k-mers with k from 1 to 8 without assuming prior knowledge about the motif, our model learns an unbiased vocabulary of TF binding.

- Unlike existing methods which only model the occurrence of the DNA sequence, our model takes the spatial effect of DNA sequence on TF binding into account. This enables the model to characterize the effect of co-factors and other auxiliary sequences in the process of transcription factor binding.

- Our method substantially outperformed existing motif-based method in predicting genetic variants reported to cause alelle-specific binding.

4.2 Future Work

The future direction of WAVE can be summarized in three categories.

- **Interactive interface**

  The first future extension of WAVE is to design a an interactive website that takes requests from users and return the scoring results for the variants re-

quested. A difficulty for this interface is the long runtime of k-mer model (usually around 3 days for one experiment). However, this can be circumvanted by pre-training k-mer model on a library of common TFs and providing scoring services only for these TFs.

- **Use WAVE predictions for other application**

  The computational nature of WAVE makes it very efficient to predict the effect of many variants on the binding of many TFs. Therefore the second possible direction is to use the prediction from WAVE for solving other interesting biological problems. For instance, previous work [25, 26] showed the possibility that a change in histone modifications is associated with change in TF binding. Therefore, one might be able to build a computational model that takes the WAVE-predicted effect of variants on the binding of a library of TFs as input, and predicts their effect on histone modifications.

- **Adopt the methodology to other high-throughput datatype**

  The third direction is to use WAVE to model the effect of variants on other high-throughput data such as ATAC-seq, histone marks ChIP-seq data, etc.

Bibliography

[1] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John P a Ioannidis, and Joel N Hirschhorn. "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." In: *Nature reviews. Genetics* 9.5 (May 2008), pp. 356–69.

[2] Teri A Manolio. "Genomewide association studies and assessment of the risk of disease." In: *The New England journal of medicine* (2010).

[3] Barbara E Stranger, Eli a Stahl, and Towfique Raj. "Progress and promise of genome-wide association studies for human complex trait genetics." In: *Genetics* 187.2 (Feb. 2011), pp. 367–83.

[4] Lucia a Hindorff, Praveen Sethupathy, Heather a Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri a Manolio. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." In: *Proceedings of the National Academy of Sciences of the United States of America* 106.23 (June 2009), pp. 9362–7.

[5] Kelly a Frazer, Sarah S Murray, Nicholas J Schork, and Eric J Topol. "Human genetic variation and its contribution to complex traits." In: *Nature reviews. Genetics* 10.4 (Apr. 2009), pp. 241–51.

[6] Lucas D Ward and Manolis Kellis. "Interpreting noncoding genetic variation in complex traits and human disease." In: *Nature biotechnology* 30.11 (Nov. 2012), pp. 1095–106.

[7] Malin C Andersen, Pär G Engström, Stuart Lithwick, David Arenillas, Per Eriksson, Boris Lenhard, Wyeth W Wasserman, and Jacob Odeberg. "In silico detection of sequence variations modifying transcriptional regulation." In: *PLoS computational biology* 4.1 (Jan. 2008), e5.

[8] Geoff Macintyre, James Bailey, Izhak Haviv, and Adam Kowalczyk. "is-rSNP: a novel technique for in silico regulatory SNP detection." In: *Bioinformatics (Oxford, England)* 26.18 (Sept. 2010), pp. i524–30.

[9] Thomas Manke, Matthias Heinig, and Martin Vingron. "Quantifying the effect of sequence variation on regulatory interactions." In: *Human mutation* 31.4 (Apr. 2010), pp. 477–83.

[10] Alberto Riva. "Large-scale computational identification of regulatory SNPs with rSNP-MAPPER." In: *BMC genomics* 13 Suppl 4.Suppl 4 (Jan. 2012), S7.

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

[12] Mingxiang Teng, Shoji Ichikawa, Leah R. Padgett, Yadong Wang, Matthew Mort, David N. Cooper, Daniel L. Koller, Tatiana Foroud, Howard J. Edenberg, Michael J. Econs, and Yunlong Liu. "Regsnps: A strategy for prioritizing regulatory single nucleotide substitutions". In: *Bioinformatics* 28.14 (2012), pp. 1879–1886.

[13] Ivan Molineris, Davide Schiavone, Fabio Rosa, Giuseppe Matullo, Valeria Poli, and Paolo Provero. "Identification of functional cis-regulatory polymorphisms in the human genome." In: *Human mutation* 34.5 (May 2013), pp. 735–42.

[14] Tatsunori Hashimoto, Richard I Sherwood, Daniel D Kang, Nisha Rajagopal, Amira A Barkal, Haoyang Zeng, Bart JM Emons, Sharanya Srinivasan, Tommi Jaakkola, and David K Gifford. In: *Under review* (2015).

[15] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12 (2011), pp. 2121–2159. arXiv: `arXiv:1103.4296v1`.

[16] Yuchun Guo, Shaun Mahony, and David K Gifford. "High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints." In: *PLoS computational biology* 8.8 (Jan. 2012), e1002638.

[17] Shaun Mahony and Panayiotis V. Benos. "STAMP: A web tool for exploring DNA-binding motif similarities". In: *Nucleic Acids Research* 35 (2007), pp. 253–258.

[18] Marco Sgarbanti, Anna L Remoli, Giulia Marsili, Barbara Ridolfi, Alessandra Borsetti, Edvige Perrotti, Roberto Orsatti, Ramona Ilari, Leonardo Sernicola, Emilia Stellacci, Barbara Ensoli, and Angela Battistini. "IRF-1 is required for full NF-kappaB transcriptional activity at the human immunodeficiency virus type 1 long terminal repeat enhancer." In: *Journal of virology* 82.7 (2008), pp. 3632–3641.

[19] Shuichi Fujioka, Jiangong Niu, Christian Schmidt, M Guido, Bailu Peng, Tadashi Uwagawa, Zhongkui Li, Douglas B Evans, James L Abbruzzese, Paul J Chiao, and Guido M Sclabas. "NF-κB and AP-1 Connection: Mechanism of NF- κ B-Dependent Regulation of AP-1 Activity". In: *Society* 24.17 (2004), pp. 7806–7819.

[20] Myriam Bartels, Aike Torben Schweda, Ursula Dreikhausen, Ronald Frank, Klaus Resch, Winfried Beil, and Mahtab Nourbakhsh. "Peptide-mediated disruption of NFkappaB/NRF interaction inhibits IL-8 gene activation by IL-1 or Helicobacter pylori." In: *Journal of immunology (Baltimore, Md. : 1950)* 179 (2007), pp. 7605–7613.

[21] R S Thomas, M J Tymms, L H McKinlay, M F Shannon, a Seth, and I Kola. "ETS1, NFkappaB and AP1 synergistically transactivate the human GM-CSF promoter." In: *Oncogene* 14 (1997), pp. 2845–2855.

[22] Joel Rozowsky, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, Nitin Bhardwaj, Mark Rubin, Michael Snyder, and Mark Gerstein. "AlleleSeq: analysis of allele-specific expression and binding in a network framework." In: *Molecular systems biology* 7.522 (Jan. 2011), p. 522.

[23] Konrad J Karczewski, Joel T Dudley, Kimberly R Kukurba, Rong Chen, Atul J Butte, Stephen B Montgomery, and Michael Snyder. "Systematic functional regulatory assessment of disease-associated variants." In: *Proceedings of the National Academy of Sciences of the United States of America* 110.23 (June 2013), pp. 9607–12.

[24] Maya Kasowski, Fabian Grubert, Christopher Heffelfinger, Manoj Hariharan, Akwasi Asabere, Sebastian M Waszak, Lukas Habegger, Joel Rozowsky, Minyi Shi, Alexander E Urban, Mi-Young Hong, Konrad J Karczewski, Wolfgang Huber, Sherman M Weissman, Mark B Gerstein, Jan O Korbel, and Michael Snyder. "Variation in transcription factor binding among humans." In: *Science (New York, N.Y.)* 328.5975 (Apr. 2010), pp. 232–5.

[25] Graham McVicker, Bryce van de Geijn, Jacob F Degner, Carolyn E Cain, Nicholas E Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K Pritchard. "Identification of genetic variants that affect histone modifications in human cells." In: *Science (New York, N.Y.)* 342.6159 (Nov. 2013), pp. 747–9.

[26] Helena Kilpinen et al. "Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription." In: *Science (New York, N.Y.)* 342.6159 (Nov. 2013), pp. 744–7.