REVEALING THE IMPACT OF SEQUENCE VARIANTS ON TRANSCRIPTION FACTOR BINDING AND GENE EXPRESSION

by

Wenqiang Shi

M.Sc., National University of Defense Technology (China), 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(BIOINFORMATICS)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2017

Abstract

Transcription factors (TFs) can bind to specific regulatory regions to control the expression of target genes. Disruption of TF binding is regarded as one of the key mechanisms by which regulatory variants could act to cause disease. However predicting the functional impact of variants on TF binding remains a major challenge for the field, standing as a key obstacle to achieving the potential of clinical genome analysis. This thesis confronts this challenge from a bioinformatics perspective and addresses two unresolved problems.

The first problem is the determination of which genetic variants alter TF binding. Only a small number of allele-specific binding (ASB) events, in which TFs preferentially bind to one of two alleles at heterozygous sites in the genome, have been determined. To study the impact of variants on TF binding, access to a large, gold standard collection of ASB events could facilitate the development of new predictive methods. In Chapter 2, we implemented a pipeline to identify ASB events from ChIP-seq data and applied it to produce one of the largest ASB datasets. We found that ASB events were associated with allelic alterations of TF motifs, chromatin accessibility and histone modifications. Using the available features, classifiers were trained to predict the impact of variants on TF binding. To improve ASB calling, Chapter 3 evaluated five statistical methods, ultimately supporting a method that pooled ChIP-seq replicates and utilized a binomial distribution to model allelic read counts.

The second problem is to determine how altered TF binding events impact the expression of target genes. In Chapter 4, we implemented regression-based models to predict gene expression changes based on altered TF binding events across 358 individuals. The models showed

predictive capacity for 19.2% of genes, and the key TF binding events in the model provided mechanistic insights as to how these regulatory variants alter gene expression.

In summary, this thesis both generated the largest, high-quality collection of ASB events, and developed algorithms to predict variant impact on TF binding and gene expression. The presented work advances the capacity of the field to interpret regulatory variants and will facilitate future clinical genome analysis.

Lay Summary

The DNA in each human cell contains the complete instructions for producing RNA and proteins that are necessary for life. Within the DNA thousands of genes are present, each with its own set of On/Off switches that allow the genes to be active at the right moments and at the right levels. Every human's DNA is slightly different, creating a rich diversity of characteristics. Some of these differences impact the activity of genes, often in subtle ways that can have impacts on health and disease. In this thesis the research explores the creation of computer methods that help identify which DNA sequence differences impact the activity of genes by altering the On/Off switches, and mathematical models that predict the gene activity based on these differences. In the long-term, the research will help us understand how each person's DNA protects them from or increases risk for health problems.

Preface

Parts of Chapter 1 have been published as a review: Mathelier, A., **Shi, W**. and Wasserman, W.W. (2015) Identification of altered *cis*-regulatory elements in human disease. Trends in genetics: TIG, 31, 67-76. The published portions from the paper include sections 1.3.1-1.3.5, 1.3.6 (updated), and 1.4.1. I wrote these sections for the original manuscript, which were subsequently revised by AM and WWW.

A version of Chapter 2 has been published: **Shi, W.**, Fornes, O., Mathelier, A. and Wasserman, W.W. (2016) Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic acids research, gkw691. With WWW and AM, I contributed to the study design and conceived the research. OF conducted the analysis for dimer transcription factors in section 2.3.3 and generated Figure 2.2B. I compiled all the data, interpreted the data with WWW and OF, and generated all figures and all tables except Figure 2.2B. I wrote the manuscript, which OF, AM, and WWW reviewed and revised.

Chapter 3 is original work in preparation for submission. With my supervisor WWW, I conceived the research and designed the study. I conducted all the analyses and generated all figures and all tables, with advice and feedback from OF and WWW. I wrote the manuscript, which OF and WWW reviewed and revised.

Chapter 4 is currently under review: **Shi, W.**, Fornes, O. and Wasserman, W.W.: Predicting the impact of altered TF binding on gene expression based on *cis*-regulatory variants. With WWW, I conceived the research and designed the study. I conducted all the analyses with WWW and OF,

and generated all figures and all tables. I wrote the manuscript, which OF and WWW reviewed and revised.

Table of contents

List of tables

List of figures

List of abbreviations

ASB: Allele-specific binding

ASE: Allele specific expression

AUPRC: Area under precision-recall curve

Bp: Base pair

ChIP-seq: Chromatin immunoprecipitation with massively parallel DNA sequencing

DHS: DNase I hypersensitivity

eQTL: Expression quantitative trait loci

FDR: False discovery rate

GWAS: Genome-wide association study

HCT: Homotypic clusters of TFBSs

IC: Information content

k-mer: DNA sequence in a fixed-length of k

Kb: 1000 base pairs

LCL: Lymphoblastoid cell line

MAF: Minor allele frequency

Mb: 10^6 base pairs

PBM: Protein binding microarrays

PFM: Position frequency matrix

PWM: Position weight matrix

QTL: Quantitative trait loci

R^2: R square

SNP: Single nucleotide polymorphism

SNV: Single nucleotide variant

TF: Transcription factor

TFBS: Transcription factor binding site

TSS: Transcription start site

WGS: Whole genome sequencing

Acknowledgements

I offer my enduring gratitude to my supervisor, Dr. Wyeth W. Wasserman, for his guidance, efforts, and supports throughout my Ph.D. study. I would like to thank my committee members, Dr. Inanc Birol, Dr. Jennifer Bryan, and Dr. Sohrab Shah, for their generous time in my committee meetings and critical questions for my research. I am grateful for the friendly academic environment of Wasserman lab. Special thanks are owed to Dora Pak for lab management, Anthony Mathelier for research advice, Oriol Fones for the help on thesis writing, Yifeng Li for advice in machine learning, Casper Shyr, Julie Chen, and all other members in Wasserman lab.

I am grateful for the funding from China Scholarship Council (four year doctoral fellowship) and Wasserman lab. I would like to thank my previous university (National University of Defense Technology) and supervisor Dr. Zhenghua Wang in China for the opportunity to study aboard.

Finally, special memory goes to my father, Runxi Shi, for his unconditional love. Special thanks are owed to my family - my mother, Ruifeng Chen, my sister, Wenjuan Shi, and my fiancée, Danyu Yao - for the support throughout my education.

Dedication

This thesis is dedicated to my family.

Chapter 1: Introduction

The diploid human genome is composed of about 6 billion DNA base pairs, which are packaged in 23 pairs of chromosomes within the nucleus of each cell. With recent advances in DNA sequencing technology, comprehensive detection of the sequence variations present in individual genomes is becoming more affordable. For instance, the Novaseq sequencing platform launched by Illumina in 2017 is able to sequence the human genome within one day at the cost of US$100. Compared with the human reference genome, a typical human individual carries about five million variants, including single nucleotide variants (SNVs), small insertions and deletions (indels), and large structural variants [1]. Various human diseases are associated with genetic variations, such as type 2 diabetes [2], multiple types of cancer [3-6], and osteoarthritis [7]. In addition, the ClinVar database has archived ~26K variants of clinical significance (as of Dec 2016) [8]. Deciphering the functional roles of these variants has been a main theme of molecular biology and medical genetics over the past decade.

1.1 The challenge of interpreting regulatory variants in the human genome

The human genome can be conceptually divided into protein coding regions and non-coding regions (DNA sequences that do not encode protein). Protein coding regions only account for a small portion of the genome (~2%), while non-coding regions, account for the remaining 98%. The non-coding portion harbors segments performing crucial functions, including regulating when and where genes are expressed (i.e. regulatory regions). To date, clinical approaches using DNA sequencing have focused on variants within protein-coding regions, which have been well characterized and are well understood relative to non-coding regions. In contrast, genome-wide association studies have identified that most disease-associated variations are situated within

non-coding regions [9, 10], especially enriched within regulatory regions [11]. In selected cases [6, 12-15], analysis of regulatory variations has led to the discovery of causal variations for genetic disorders. However, our understanding of these regulatory variants remains incomplete, making it difficult to predict the impact of individual variants on gene expression and disease. Hence, in the whole genome sequencing (WGS) era, there is an imperative need for informatics methods to predict both the locations and specific functions of regulatory variants.

1.2 Basics of gene regulation and transcription factor binding

Genes contain the instructions to synthesize proteins and functional RNAs that participate in the physiological activities of cells. These instructions direct gene expression at multiple levels, including: 1) transcriptional regulation, which controls when and to what extent DNA is transcribed to RNA; 2) post transcriptional regulation, which controls how RNAs are processed and transported; and 3) translational regulation, which controls how proteins are synthesized from mRNA. Additional gene expression-related mechanisms regulate the stability and activity of RNA or proteins within cells, but these will not be further addressed in this thesis. Each mechanism is important, but much of the global research effort has focused on the control of transcription, both because experimental methods are well developed and because transcription controls the inputs to downstream regulatory processes.

Transcriptional initiation is coordinated by transcription factors (TFs), proteins that are involved in the production of RNA (including proteins that do not bind to DNA in a sequence specific manner). Gene transcription begins with sequential binding events of various TFs to promoters, DNA segments from which RNA transcripts will be initiated. The assembled TFs at a promoter

ultimately recruit RNA polymerase to initiate transcription (Figure 1.1). TF binding at regions distal to promoters (e.g. enhancers or silencers) can influence the transcription rate, for instance, by interactions with promoters in 3D space [16]. Spatial and temporal combinations of TFs provide the means for cells to exquisitely control gene expression at different cell developmental stages or in response to dynamic environmental conditions.

TFs are the main regulators in transcriptional regulation. Subset of TFs contain DNA-binding domains, which recognize and bind to DNA in a sequence-specific manner (usually 6-19 bp in human). The DNA sequences bound by TFs are called TF binding sites (TFBSs). Beyond TFBS, TF binding is strongly influenced by chromatin accessibility, epigenetic marks, and chromatin 3D architecture [17]. Various experimental technologies and computational approaches have been developed to identify TF binding activity, expanding our knowledge on transcriptional regulation.

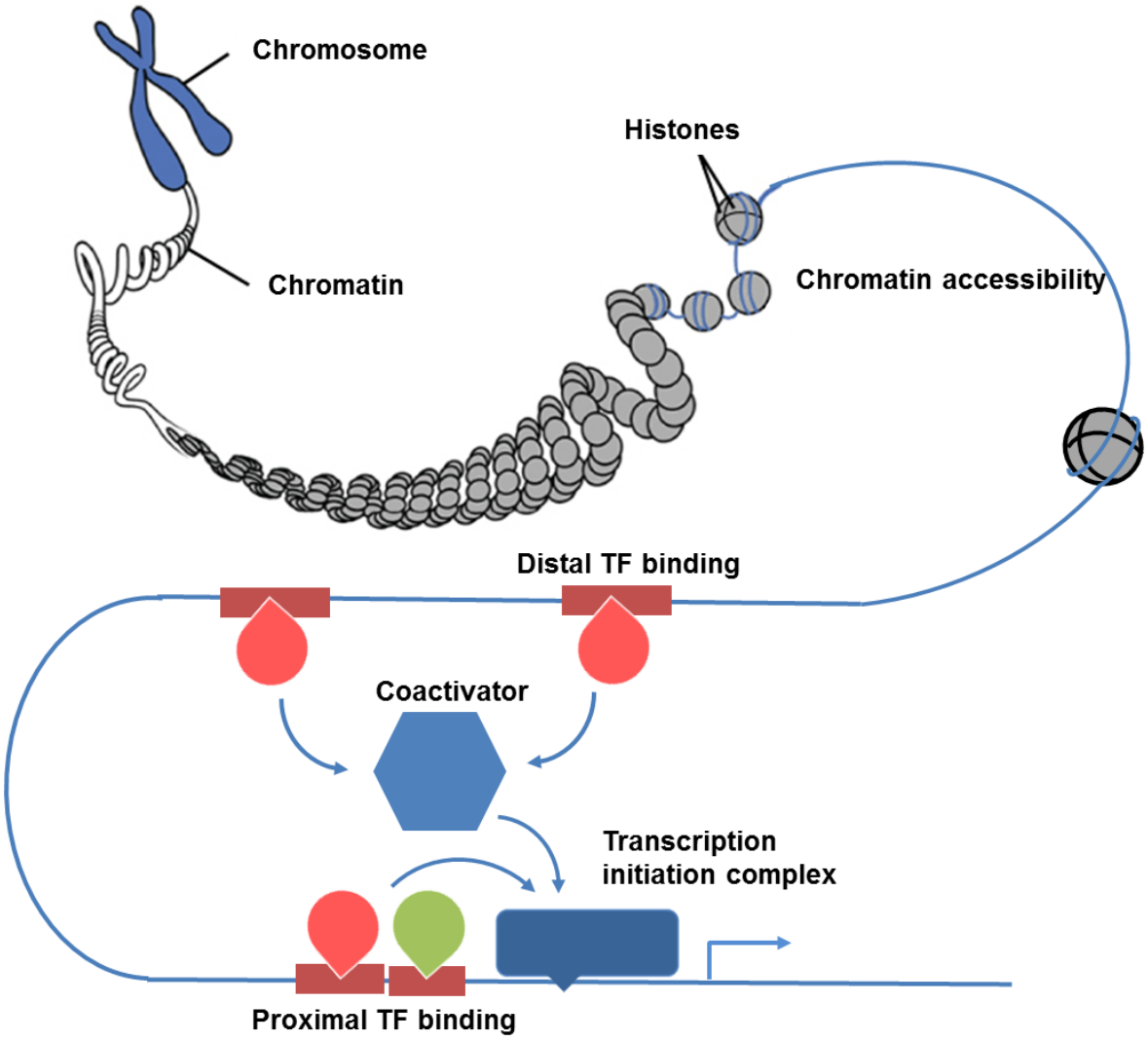


Figure 1.1 Transcriptional regulation

Gene transcription is initiated and coordinated by a series of TF binding events at promoter and distal regulatory regions. TF binding is a complex biochemical mechanism determined not only by DNA sequences, but also influenced by chromatin states, such as chromatin accessibility and histone modifications. (Figure modified from https://en.wikipedia.org/wiki/File:0321_DNA_Macrostructure.jpg licensed under the Creative Commons Attribution 4.0 International license)

1.2.1 Experimental detection of TF binding

TF binding can be detected by multiple experimental techniques in a high throughput manner, including protein binding microarrays (PBM) [18], bacterial one-hybrid screening [19], high throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) [20], and chromatin immunoprecipitation followed by massively parallel DNA sequencing (ChIP-seq) [21]. The first three approaches detect TF binding *in vitro*. For instance, PBM experiments can test the binding affinity of the target TF against all possible DNA sequences of a given length (e.g. 10bp), significantly expanding potential TF binding sequences not present in the genomes of living organisms. However, *in vitro* binding assays only measure the contribution of DNA sequences to TF binding, and cannot account for the *in vivo* properties such as chromatin accessibility, epigenetic marks and partner TFs.

The development of ChIP-seq technology allows researchers to identify where a TF of interest binds to the DNA at genome-scale *in vivo*. In a ChIP-seq experiment [21], TF-DNA complexes are firstly cross-linked and then the chromatin in the nucleus is sheared to DNA fragments of several hundred base pairs. The targeted cross-linked TF-DNA complexes are captured and enriched by a TF-specific antibody in the immunoprecipitation process. Finally, the DNA fragments in the enriched complexes are sequenced. For a typical ChIP-seq experiment studying human sequence-specific DNA binding TFs (e.g. CTCF), 20 million sequenced reads are recommended for downstream analysis [22]. Proteins which interact with DNA more broadly (e.g. RNA Pol II or histone variants) require a larger number of reads (e.g. 60 million).

After sequencing, ChIP-seq experiments require downstream bioinformatics analysis to call potential TF-bound regions. Reads of ChIP-seq experiments are firstly mapped to a reference genome using read aligners, e.g. BWA [23], Bowtie [24], or Novoalign (http://novocraft.com/). Mapped reads are expected to be enriched in putative TF-bound regions, which are visualized as peaks of mapped reads across a chromosome. Peak-calling algorithms [25, 26] are used to identify putative TF-bound regions. Typical human sequence-specific TFs exhibit on the order of thousands of peaks, and the resolution of the width of called peaks is usually hundreds of base pairs. The real TF binding sites (6-19bp) are expected to be enriched around peakMax positions (the position with the highest number of mapped reads within a peak) for high quality ChIP-seq experiments [27]. More advanced ChIP-based techniques are emerging which provide higher resolution for TF-bound regions, such as ChIP-exo [28, 29].

The ENCODE project [30] provides over one thousand datasets to investigate the functional elements in the human genome, including ChIP-seq data for histone modifications and hundreds of TFs, DNase-seq data for DNase I hypersensitivity (DHS) sites (detecting accessible regions within the genome), and RNA-seq for gene expression. The ENCODE project found that up to 19.4% of the genome is marked as regions of transcription factor binding or DHS at least in one cell type [30]. This suggests that a great portion of non-coding regions are involved in gene regulation, although such involvement may be limited to specific developmental or environmental contexts. These rich datasets create new opportunities for bioinformatics approaches to improve the prediction of TF binding.

1.2.2 Computational prediction of TF binding

A subset of TFs bind to DNA in a sequence specific manner, and their binding sequences can be identified through any of the aforementioned experiments. Based on experimentally validated bound sequences, TF binding preferences can be computationally modeled using various approaches described below.

1.2.2.1 Classic TF binding model: position weight matrix

The most widely used predictive model for TF binding is the position weight matrix (PWM) (Figure 1.2) [31]. To build a PWM for a given TF, experimentally collected TFBSs are aligned to derive a position frequency matrix (PFM), which summarises the frequency of each nucleotide at each binding site position (Figure 1.2B). The PFM is then normalized per binding position and transformed to a PWM representing the binding preference at each position relative the background nucleotide distribution (Figure 1.2C). Given a candidate site, the binding score of the site is the sum of PWM values corresponding to the sequence nucleotide at each binding position. Sites with PWM scores above a certain threshold are regarded as predicted TFBSs (Figure 1.2D). PWM scores of TFBSs have been shown to correlate with the binding energy of TF-DNA interactions [31]. Compiled PWMs can be obtained from multiple open access databases, including JASPAR [32], HoCoMoCo [33], and CIS-BP [34].

PWMs assume that each TFBS position independently contributes to the overall TF binding [17]. This assumption simplifies the procedure of PWM score calculation, but it neglects dependencies between neighbor nucleotides within the TFBS. To address this issue, nucleotide dependency within the TFBS has been accounted in advanced models, such as ones using hidden Markov

models [35] or Bayesian networks [36]. While the advanced models achieved modest improvements in some cases, overall, the PWM remains a simple and powerful method for predicting TF binding.

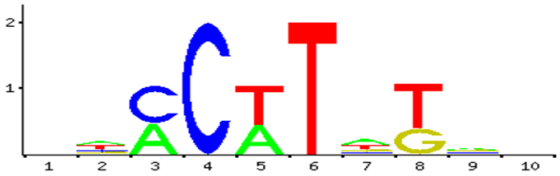
(A) Known binding sequences

Seq 1	A	C	C	**C**	T	**T**	A	T	T	C
Seq 2	C	T	C	**C**	A	**T**	C	T	C	A
…	…				…				…	
Seq 58	T	G	A	**C**	T	**T**	T	G	G	T
Seq 59	G	A	A	**C**	A	**T**	G	G	A	G

(B) Position frequency matrix

A	16	20	26	0	25	0	24	0	17	15
C	15	18	33	59	0	0	5	0	14	15
G	14	5	0	0	0	0	12	22	14	15
T	14	16	0	0	34	59	18	37	14	14

(C) Position weight matrix logo (PWM)



(D) Predict TFBS in new sequence

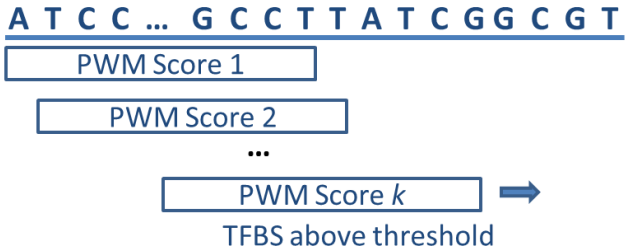


Figure 1.2 Position weight matrix and motif scoring

(A) Align known binding sequences of the investigated TF. (B) The frequency of each nucleotide at each binding site position can be summarized into a PFM. (C) A PFM can be normalized and log-transformed to a PWM, which

can be represented as a motif logo. In the logo, large-size nucleotides indicate the key positions within the TF binding site. (D) Scan DNA sequence with PWM to identify potential TFBSs. The PWM is aligned to the sequence, and the score of matched DNA sequences (of same length as the PWM) is the sum of the corresponding nucleotide values at each column of the PWM. Only the sites with a score reaching the predefined threshold are regarded as predicted TFBSs.

1.2.2.2 *k*-mer based approaches

Instead of the traditional PWM approach, TF binding preference can also be modeled by a set of DNA sequences in a fixed-length of k (referred as *k*-mers). In *k*-mer approaches, the occurrences of every *k*-mer are input features for the model to distinguish between TF-bound and background regions [37, 38]. On top of the core motif of TF, the *k*-mer approach can capture other sequence patterns like redundant motifs or motifs of partner TFs, enhancing predictive performance [37, 39]. In addition, *k*-mer approaches are able to predict regulatory regions like DHS and enhancers, which are composed of multiple TFBSs [38, 40].

1.2.2.3 Deep convolutional neural networks methods

Deep convolutional neural networks are a special kind of multiple-layer neural network, which can model high-level abstraction and non-linear relationships in rich datasets [41]. They have been successfully applied to image and speech recognition, showing superior performance than previous methods [41, 42]. Deep convolutional neural networks usually require large amounts of training data to avoid over-fitting. Recent advances in sequencing technologies, including ChIP-seq and DNase-seq, have identified thousands of TF-bound and open chromatin regions across the genome. Multiple deep learning methods have trained models based on these datasets to predict TF-bound regions, including DeepBind [43], DeepSEA [44] and DanQ [45]. Similar to

the task of image recognition, these methods usually treat the genomic sequence (up to 1000bp) as a one-dimension image composed of four colors. Then, the TF binding prediction problem is transformed to an image classification task. As expected, deep learning based models outperform most traditional models [43, 44].

1.3 Assessing the impact of variants on TF-DNA interactions

After identifying TF-bound regions, the next challenge is to predict if variations within these regions are likely to alter TF binding. TF-DNA interactions arise from the interplay between DNA sequence motifs, chromatin accessibility, epigenetic marks, and interactions with other partner TFs. To develop methods to predict the subset of variants most likely to disrupt TF binding, one requires a set of reliable data and should consider multiple aspects of TF-DNA interactions.

1.3.1 Collecting reliable reference data sets

To access data that enables the creation of predictive models for the functional impact of variations within TFBS, it would be convenient to utilize the extensive body of ChIP-seq experimental data. A subset of cell lines have been studied by both the ENCODE and the 1000 Genomes projects, providing the community with TF binding, epigenetic, and genotyping data from the same cellular context. The union of these data allows for in-depth analyses of the impact of variations within TFBSs [46-48].

Allele-specific binding (ASB) events, in which a TF preferentially binds to one allele at heterozygous sites, provide a valuable capacity to study the impact of single nucleotide

variations on TF binding within the same cellular environment [47, 49] (Figure 1.3). Multiple pipelines are now available to detect ASB events based on TF ChIP-seq experiments and WGS data from the same cell [49-52] (Figure 1.4). In an ASB pipeline, using a single reference genome for the mapping of ChIP-seq reads will introduce reference bias; ChIP-seq reads with non-reference alleles of heterozygous sites are less likely to map to the correct site compared with the reads matching the reference allele [49, 53]. To address this issue, a personalized reference genome approach allows equal success in mapping ChIP-seq reads from both alleles in an individual. The significance of ASB events can be assessed using a binomial or beta-binomial test applied to the number of mapped reads for each of the two alleles [47, 49]. Detectable ASB events only represent a small portion of any ChIP-seq dataset—usually less than 1% of all peaks [49]. ASB events are enriched for disease-associated SNPs and when situated within 100bp of promoter TSSs are strongly associated with gene expression alteration [47].

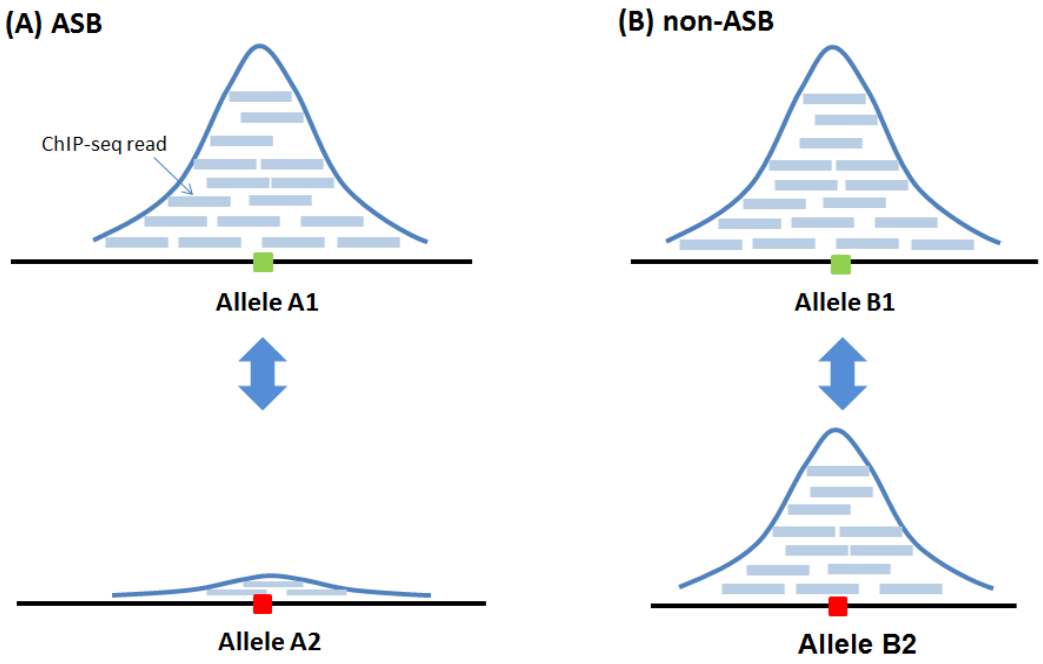


Figure 1.3 Allele specific binding and non-ASB events

(A) ASB events can be derived from heterozygous loci where there is a strong preference for ChIP-seq reads to map to one allele. (B) Non-ASB events harbour reads mapping to both alleles similarly. It could be explained by the TF equally recognizing the two alleles or the variation not disrupting the active binding site.

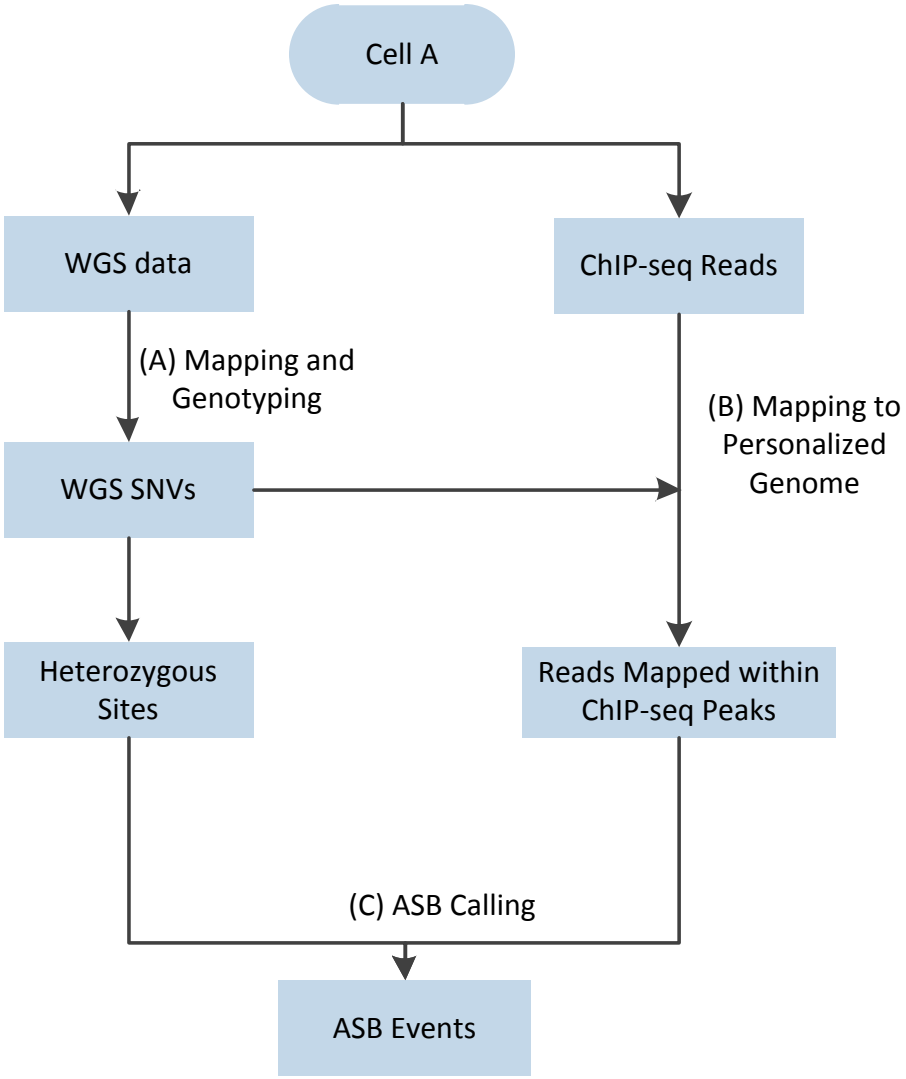


Figure 1.4 Bioinformatics pipeline to identify ASB events

This pipeline takes as input both ChIP-Seq and WGS data from the same cell. (A) The first step consists of mapping the WGS data to the reference genome, and calling the corresponding genotype. If genotype data are available, this step can be skipped. (B) The called genotypes are then used to create the associated personalized genome to map ChIP-Seq reads. (C) The final step is to extract the number of mapped reads for each allele at known heterozygous sites within ChIP-seq peaks, and then an ASB event is called if the number of mapped ChIP-seq reads on one allele is significantly different from the other allele.

1.3.2 Interpreting TF binding alteration

Differential TF binding has been analyzed across individuals (in either tissue samples or cell strains) or at heterozygous sites in individual cell lines. Motif altering SNVs account for a subset of the observed binding differences [46, 54] (Figure 1.5AB). Alleles more similar to the consensus motif preferentially show more elevated binding [47, 48, 55]. In addition, altered motifs proximal to an experimental ChIP-seq peak max position are more associated with differential binding [54]. Most ASB do not overlap with the motif of ChIP'ed TF (the targeted TF of the ChIP-seq experiment), however, indicating that other mechanisms account for an important portion of observed ASB events [47, 48]. The presence (or absence) of cofactors (i.e. TFs acting cooperatively) binding nearby could account for a subset of these events (Figure 1.5AC) [54, 56]. For instance, NF-kB binding differences between individuals are correlated with SNVs altering TFBSs of its cofactors [56]. Overall, the variation within the ChIP'ed TF motif or a cofactor motif can lead to altered TF-DNA interaction and explains a significant proportion (e.g. 37.5% in [48]) of ASB events. The relative impact of other potential contributory mechanisms remains to be evaluated.

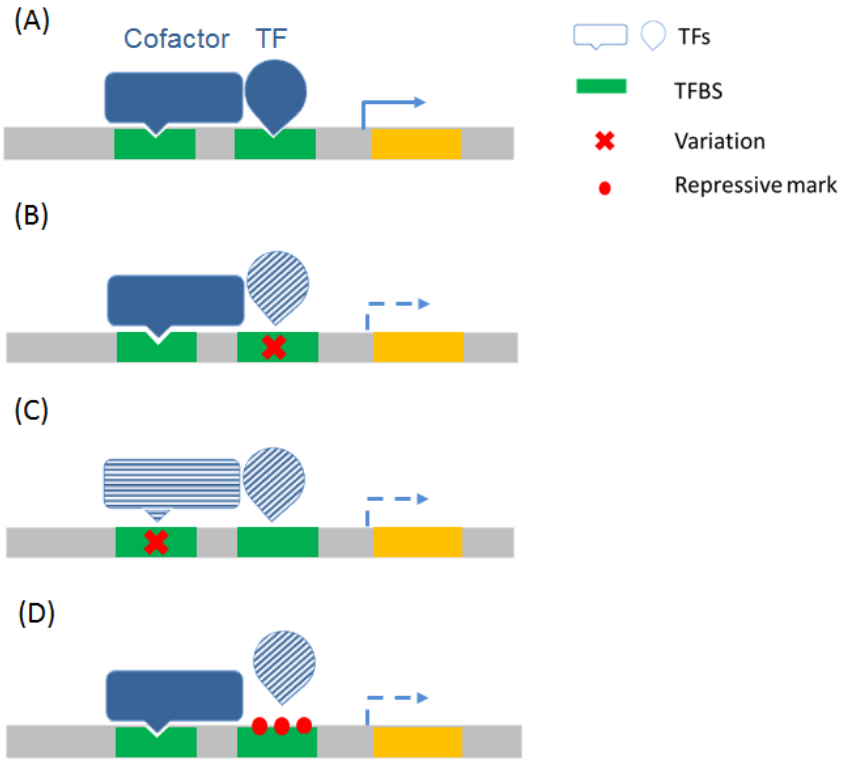


Figure 1.5 Schematic view of TF binding alteration

(**A**) TF binding events in a "normal" environment. Two TFs bind to their TFBSs and stabilize each other's binding. (**B**) A variant in the TFBS disrupts the binding of the TF to DNA (light shading). (**C**) A variant in one TFBS disrupts the binding of both TFs to DNA. (**D**) Modification of the epigenetic environment represses the binding of the TF to DNA.

1.3.3 Chromatin marks and TF binding

TFs can show different binding preferences in the context of specific open chromatin and histone modifications, and this preference has been used to inform TFBS prediction [57]. However, the dependence between chromatin marks and TF binding is not always clear [58]. There are subsets of epigenetic marks linked to increased TF binding (active marks), and subsets linked to reduced binding (repressive marks) [59] (Figure 1.5D). From the opposite perspective, studies show the

contribution of TF binding to the specification of histone modifications [26, 60]. Moreover, SNPs are enriched in regions with variable epigenetic marks between individuals when compared to invariant regions [55]. Taken together, it is possible that some TF binding events depend on specific chromatin marks, whereas other TFs do not (such as the so-called pioneer factors, see [61]). Still, attributing differential TF binding to epigenetic alterations, cofactor binding, or canonical TFBS disruption, remains a challenge.

1.3.4 TFBS redundancy

Studies have shown that the presence of redundant binding sites in *cis*-regulatory regions helps maintain the pattern and level of expression, even in the event of sequence alterations [48, 62, 63]. As reviewed in [63], TFBSs can be considered as incremental inputs to the transcriptional regulation with varying the degree of potency and redundancy. TFBS redundancy can be considered as a buffering mechanism, through which a disrupted TFBS can be compensated by another nearby binding site in the same *cis*-regulatory region [48, 62] . It may therefore be important to consider redundant TFBSs when interpreting the functional impact of variants on gene regulation.

1.3.5 TFBS conservation

Sequence conservation can inform the interpretation of variations within *cis*-regulatory elements (as it does for protein-coding sequence). Linking TFBS sequence conservation and functionality is not trivial. For instance, an early experimental study of regulatory elements in human and mouse estimates that about 32-40% of human functional regions are not functional in mouse [64]. Moreover, TF bound regions identified in large-scale ChIP-seq experiments show limited

conservation across species, with only 10-22% of peaks conserved [65]. While many ChIP-identified regions are not conserved, the ones overlapping evolutionarily conserved sequences are associated with higher rates of functional roles [66, 67]. Moreover, genomic sequences under lineage specific selection have been used to filter *cis*-regulatory variations in cancer WGS analyses [68].

1.3.6 Computational tools to predict variation impact on TF binding

Tools have been developed to predict the impact of variants within TFBSs by integrating both experimentally derived and sequence-based features. Early methods focused mainly on the use of TF binding profiles to evaluate the impact of variants on TF-DNA binding interaction strength. RAVEN [69] uses phylogenetic footprinting information along with PWM score differences between reference and alternative alleles. Tools like is-rSNP [70] and regSNP [71] evaluate TFBS-altering events by the significance of PWM score change. The sTRAP program [72] and BayesPI-BAR [73] assess the binding affinity difference between wild-type and mutated sequences for TFBSs by using a biophysical model with available TF binding profiles. Advanced TF binding models (e.g. DeepSEA[44] and gkmSVM [40]) capture more complex patterns within broader TF regions beyond core TFBSs. Binding score differences of these models can be predictive for the variation impact on TF binding.

Other approaches incorporate epigenetic data along with TF binding profiles to assess the impact of regulatory variants in genome-wide association studies (GWAS). GWAS3D [9] integrates chromosomal capture information along with epigenetic marks, binding affinity impacts based on scores from PWMs, and conservation to prioritize regulatory variants. The impact of variants

on TF-DNA affinity is assessed by comparing the log-odds binding probabilities between the reference and alternative alleles to a null empirical distribution. Funseq2 [74] prioritizes regulatory variants in cancer by incorporating the alteration of TFBSs, inter- and intra-species conservation, and gene networks, and the weights of non-coding features is estimated by mutation patterns across the 1000 Genomes SNPs.

To predict the variant impact on TF binding, current progress is hampered by the limited number of variants which are known to alter TF binding. For instance, to verify BayesPI-BAR models, Wang *et al.* collected 67 variants which have been experimentally validated to alter TF binding [73]. This set is too small to train discriminative models which can learn the boundary between disruptive or non-disruptive variants directly. Alternatively, existing methods (e.g. BayesPI-BAR [73], is-rSNP [70]) score the binding potential of the two alleles based on DNA sequence and then evaluate the amount of difference. However, it is challenging for existing methods to justify a boundary (threshold) for disruptive variants. Recently, ASB events are emerging as gold-standard datasets for altered TF binding events and can be compiled in high throughput way [47, 49], representing future training data for new discriminative models.

1.4 Assessing the impact of *cis*-regulatory variants on gene expression

For the regulation of RNA transcript initiation, *cis*-regulatory regions include (but are not limited to) the non-coding DNA regions containing TFBSs, comprising promoters, enhancers and silencers. As introduced above, promoters are the regions from which RNA polymerase initiates production of RNA. Enhancers and silencers are distal *cis*-regulatory regions where TFs bind to fine-tune gene expression levels or patterns. Enhancers can be far from promoters in the

sequence space (up to 1Mb in some cases), but they can be close to a promoter region in the 3D space through chromatin interactions. Sequence variants can disrupt the function of *cis*-regulatory regions, impacting the expression of the regulated genes. To predict the impact of *cis*-regulatory variants on gene expression one needs to: 1) identify the *cis*-regulatory regions; 2) associate the *cis*-regulatory regions to the genes they regulate; and 3) evaluate the impact of *cis*-regulatory variants on TF binding and gene expression.

1.4.1 Identifying *cis*-regulatory regions in the human genome

To infer the functional roles of *cis*-regulatory variants, a first step is to identify the locations of *cis*-regulatory regions for each gene. Although a comprehensive inventory of functional sequences and their activities across all cells and conditions is prohibitive, researchers have accumulated data to identify vast numbers of *cis*-regulatory regions. Table 1.1 succinctly introduces collections of data that are of particular value for the detection of *cis*-regulatory regions.

Features	Description
TF binding	TF binding sites are the core elements of *cis*-regulatory regions [75] and TF binding events in regulatory regions would control the expression of the targeted gene [17].
Histone modifications	Post-translational modifications at the N-terminal tail of histone proteins. Histone modifications affect the overall chromatin structure and are associated with *cis*-regulatory regions [30, 76].
Nucleosome	Basic organizational unit of eukaryotic chromatin composed of an octamer of histone protein cores and a segment of DNA. Nucleosomes are usually depleted at promoters and enhancers [77, 78].
Open chromatin	Highly accessible regions for TFs and other proteins. Open chromatin regions are usually associated with active genome activity [79-81].
DNA methylation	Various roles in gene regulation and in different cell contexts. Methylation at promoter regions is usually associated with gene silencing [82].
Chromatin conformation	Chromosomes are compacted in the nucleus of cells. Regions far-apart in the genome sequence can be close in the 3D space [83]. Chromatin conformation analyses identify genomic interactions in cells [16, 84].
Conservation	Computation of the amount of conserved nucleotides across species through genome alignments. Conserved non-coding genomic regions are more likely to be functional [74, 85].
Nucleotide sequence properties	The nucleotide composition of the genome. For instance, G+C content helps to identify nucleosome positioning [86], and CpG islands are over-represented in promoters [87].

Table 1.1 Examples of features used for the identification of *cis*-regulatory regions

1.4.1.1 Properties and experimental data of *cis*-regulatory regions

The activity of *cis*-regulatory regions is controlled through a complex interplay between epigenomic modifications, conformation of the chromatin, and binding of TFs (Table 1.1). While highly valuable, the range of histone marks and TF binding events that can be experimentally profiled remains limited by a number of high-quality specific antibodies. Active *cis*-regulatory regions are usually associated with open chromatin, which can be identified through DNase I hypersensitivity, FAIRE-seq [88], or ATAC-seq [89] experiments. In addition to experimentally derived data, one can use genomic sequence characteristics to define *cis*-regulatory regions, such as conservation (see [90] for a review) or dinucleotide composition for enhancer prediction [91].

Experimental methods can now identify active *cis*-regulatory regions in bulk. For instance, the self-transcribing active regulatory region sequencing technology (STARR-seq) allows for the identification of active enhancers by assaying millions of candidate DNA sequences [92]. The FANTOM5 consortium [93, 94] screened RNA from hundreds of mammalian samples and cells lines for active promoters and enhancers using the cap analysis of gene expression (CAGE) technique [95]. These compilations provide a rich source of training data for new informatics methods.

1.4.1.2 Identifying *cis*-regulatory regions through machine learning

Drawing upon the growing body of genome-scale datasets, new machine learning-based methods have emerged for the annotation of *cis*-regulatory regions. Unsupervised models (in which observed data properties inform the classification of genome segments into groups) segment the genome into segments [96], of which some may be highly enriched for annotated regions such as

promoters or enhancers. For instance, the ENCODE project annotated the human genome into seven chromatin states based on histone marks, open chromatin marks, CTCF binding, and RNA Pol2 binding signals [96]. Experimental validation showed that only 26% of the predicted enhancers from the ENCODE project had regulatory activity in the targeted cell lines [97], suggesting the limited prediction power of the selected training marks for *cis*-regulatory activities. Meanwhile, supervised methods take advantage of annotated regions labelled by experimental data (e.g. enhancers from the FANTOM5 project) to train predictive algorithms [98]. As more experimentally derived data sets become available for training, supervised approaches will become increasingly powerful for delineating the locations of *cis*-regulatory regions.

1.4.2 Associating regulatory regions to target genes

Distal *cis*-regulatory regions (e.g. enhancer) regulate the target gene through chromatin loops, which can be detected by chromatin conformation capture-based technologies, such as 3C, 4C, 5C, Hi-C, and ChIA-PET [99-103]. For instance, Hi-C data revealed that chromosomes are spatially partitioned into contact domains, within which chromosome regions interact more frequently than the rest of the chromosome regions. Chromatin loops connected with promoters have been found to link to known enhancers [83].

The target of *cis*-regulatory regions can also be inferred computationally. For instance, Funseq2 considers two histone modifications and DNA methylation as activity markers of *cis*-regulatory regions. The method links *cis*-regulatory regions to genes if their activities correlate with the expression levels of the gene across 20 tissues [104]. Moreover, enhancer and promoter activities

can be directly measured by CAGE technology [95]. The FANTOM5 project associates enhancers to promoters based on their CAGE activity correlation across ~800 cell types and tissue samples [94]. However, the inferred associations do not guarantee real chromatin interactions or regulatory relationships. Alternatively, enhancer–promoter interactions in Hi-C data can be predicted based on genomic features, such as TF binding, DHS, epigenetic marks and gene expression [105].

1.4.3 Approaches to evaluate the impact of non-coding variants on gene expression

High-throughput sequencing technologies have greatly improved our understanding of transcriptional regulation in multiple aspects, including sequence variation, gene expression levels, TF binding, active *cis*-regulatory regions and chromatin interactions. However, interpreting the functional roles of *cis*-regulatory variants is still challenging, both in computational and experimental scenarios.

1.4.3.1 Experimental assessment of the impact of variants on gene expression

Various experimental techniques have been developed to evaluate the impact of variants on gene expression. In massively parallel reporter assays (MPRAs) [106-108], mutated *cis*-regulatory regions are synthesized into constructs with reporter genes. The relative expression of the reporter gene from each construct provides a quantitative measure of the impact of the introduced mutations [109]. In one MPRA study, 4.6 million nucleotides were tested in 15,000 putative *cis*-regulatory regions, revealing key mutations supported by regulatory motifs and evolutionary conservation [110]. Recently, the new genome editing method, CRISPR/Cas9, has been used to precisely introduce mutations *in vivo* [111]. For instance, mutations introduced by

the CRISPR/Cas9 system in the CArG box of a gene promoter greatly decreased the expression of that gene in mice, demonstrating that CRISPR/Cas9 is an efficient approach to test the role of individual variants [112]. High-throughput CRISPR-Cas9 approaches have been developed to screen the impact of mutations across *cis*-regulatory regions of target genes, representing an alternative approach to MPRA [113]. Taken together, current approaches are able to test the impact of mutations on gene expression, but inferring their molecular mechanisms still needs downstream computational analysis.

1.4.3.2 Computational approaches to identify the impact of non-coding variants on gene expression

WGS sequencing data has revealed that a typical individual's genome harbors about five million variants [1]. The functional impacts of these variants are largely unknown, especially for non-coding variants. A powerful approach to detect the impact of non-coding variants is the quantitative trait loci (QTL) test, which can identify SNPs associated with certain measurable molecular traits across multiple individuals, such as gene expression [114, 115], DHS [116] and DNA methylation [117]. Taking expression QTLs (eQTLs) for example, a QLT test is conventionally formulated as a linear regression model

$$\boldsymbol{Y}_i \sim \beta_0 + \beta_k \boldsymbol{X}_k + \epsilon$$

where $\boldsymbol{Y}_i$ represents the expression vector of gene *i* across testing individuals, $\boldsymbol{X}_k$ is the genotype vector of SNP_k (encoded as the number of minor alleles) across testing individuals, β_0 is a constant indicating the mean gene expression, β_k is the effect size of SNP_k towards gene expression, and ϵ is the error term following normal distribution with zero mean and constant variance [118, 119]. The significance of the association can be obtained by testing the null

hypothesis that β_k is zero. eQTL studies usually conduct thousands of tests between genes and SNPs, and multiple testing correction is required to control the false positives [120].

Expression QTLs (eQTLs) are found to cluster around transcription start sites, and are enriched in *cis*-regulatory elements (e.g. TF ChIP-seq peaks, DHS regions, promoters, and enhancers) [121], supporting the contribution of *cis*-regulatory variants to gene expression. In addition, QTLs of regulatory traits showed a close relationship with eQTLs. For instance, a portion (16%) of DHS QTLs (dsQTLs) are associated with the expression of nearby genes, and up to 55% of eQTLs are estimated to be dsQTLs [116]. Lastly, dsQTL showed 3.6-fold enrichment in TF binding footprints, suggesting the contribution of TF binding to chromatin accessibility [116].

In QTL studies, it is still a challenge to infer causality of variants and their functional roles. Identified QTLs only indicate the existence of causal variants which can be in high linkage disequilibrium with the identified marker [122]. In addition, there might be multiple causal variants within the same linkage disequilibrium block [122]. Moreover, QTL approach often lacks sufficient statistical power to detect the impact of rare variants [114], which are responsible for many rare family genetic disorders [123].

Gene expression can be quantitatively predicted based on multiple local variants following multiple regression models [124, 125]. As the number of features can be larger than the size of training samples, penalized regression models have been widely used in model training [124-126]. The performance of such models can be improved by assigning greater weights to variants overlapping with certain genomic features, such as TF binding, untranslated regions, etc. [125].

These functional genomic features can also be integrated as priors in a Bayesian framework [127]. However, these models still focus on the relationship between genotype and gene expression, and the functional mechanism of identified variants remains elusive.

1.5 Thesis overview and objectives

Understanding the role of genetic variants in human disease is a fundamental question of medical genetics. Current clinical analyses focus on variants within protein coding regions, as interpreting the functional role of non-coding variants remains a challenge. However, the majority of disease-related variants identified in genome-wide association studies are located within non-coding regions, especially enriched in regulatory regions related to TF binding, DHS or histone modifications [11]. Disruption of TF binding has been regarded as the major mechanism for regulatory variants [128, 129]. Diseases and phenotypes caused by disrupted TF binding are being identified in experiments [6, 7, 15, 130, 131]. However it is still challenging to predict the functional impact of regulatory variants on TF binding and gene expression. My thesis addresses this challenge using computational approaches (Table 1.2).

Chapter 2 introduced novel datasets and a new framework to understand the impact of regulatory variants on TF binding. When developing methods for regulatory variants analysis, it is fundamental to compile reliable examples of altered TF binding. However, only a few hundred of variants had been experimentally validated to alter TF binding [70, 73]; this was insufficient for statistical modeling of any TF. We improved the predictions by expanding the set of ASB events, which are *bona fide* TF binding alterations and compare the impact of two alleles at heterozygous sites within the same cellular context. We developed a pipeline to extract ASB

events from ENCODE ChIP-seq data sets. We found that ASB events were frequently associated with motif alterations of the ChIP'ed TF and potential partner TFs, and allelic differences of DHS and histone modifications. Classifiers were trained based on ASB datasets to predict the impact of variants. They showed a comparable performance to other state-of-the-art algorithms.

In Chapter 3, we sought to evaluate different methods for ASB calling. Chapter 2 used the traditional binomial test to call ASB events based on the number of mapped reads on two alleles. New ASB calling approaches are emerging considering two characteristics of ChIP-seq datasets: 1) greater variance in allelic read distribution than expected in the binomial distribution [132]; and 2) variance between ChIP-seq replicates [133, 134]. In Chapter 3, we benchmarked five ASB calling methods on the compiled ASB data and identified the most appropriate method based on the allelic DHS data.

Chapter 4 addressed the key unanswered question of how altered TF binding impact gene expression. Current models [124, 125] used SNPs within the gene body and flanking regions (±1Mb) to predict the variation of gene expression. Due to the linkage between SNPs, the specific functional roles of individual SNPs are hard to distinguish. To address this problem, we built a quantitative model for the expression of each gene based on altered TF binding events in regulatory regions. Our models showed an acceptable performance for 19.2% of genes ($R^2 >$ 0.05). Alteration of DHS, and RUNX3 and CTCF binding were the most frequently selected features. Though our models showed a comparable performance to existing SNP-based models, they provide a broader insight into the mechanisms by which non-coding variants impact gene expression.

Taken together, the three components of this thesis addressed the challenge of interpreting the impact of non-coding variants on transcriptional regulation. We developed pipelines and statistical models to call ASB events, and compiled ASB events in 45 ChIP-seq experiments. We built quantitative models to predict the impact of variants on TF binding, and their subsequent impact on gene expression, providing mechanistic insights into the role of regulatory variants.

Chapter	Objective
2	Interpret the impact of SNVs on TF binding based on allele specific binding events
3	Evaluate statistical models to call ASB events
4	Predict the impact of altered TF binding on gene expression

Table 1.2 List of research objectives of this thesis

Chapter 2: Evaluating the impact of single nucleotide variants on transcription factor binding

High quality data are important for statistical analysis and machine learning algorithms. In order to assess the impact of sequence variant on TF binding, this chapter compiled sets of reliable cases in which a subtle variation has a quantitative impact on TF binding. These datasets revealed key features for altered TF binding and enabled us to train predictive models for the impact of sequence variant on TF binding.

2.1 Introduction

With recent advances in DNA sequencing technology, comprehensive analysis of sequence variants in individual genomes is possible for the first time. The technology has enabled genetics researchers to systematically seek variations that contribute to disease phenotype. Up to now, clinical approaches using DNA sequencing have focused on about 2% of the human genome containing protein-coding exons. In contrast, most disease associated variants arising from genome-wide association studies are situated within non-coding regions [9]. These regions are enriched with transcription factor (TF) binding sites (TFBSs) [68], critical sequences for the regulation of gene expression. Thus, there is a pressing need to predict the impact of genetic variations on TF binding.

The prediction of which DNA sequence alterations will alter TF binding is a long-standing challenge in bioinformatics. Progress is hampered by the limited number of reliable data sets for TF binding disruption. Although thousands of expression quantitative trait loci have been

identified, they are not suitable for the study of TF binding alteration because TF binding information is not available. Only a few hundreds of naturally occurring variations have been experimentally validated to alter the binding of TFs, with low depth for any specific TF [69, 70]. Thus, current studies cannot directly train a model on true alteration data. Instead, existing methods score the binding potential of the two alleles based on DNA sequence and then quantify the difference, with examples including is-rSNP [70], BayesPI-BAR [73], and deltaSVM [40]. However, many TF binding alterations do not arise from genetic difference within the TFBSs, as other influences can contribute, such as epigenetic variation and disrupted binding of cooperative TFs [46]. The lack of experimentally determined disruption data makes it difficult to capture multiple defining properties of disrupted TFBSs.

The availability of large-scale data obtained through the chromatin immunoprecipitation followed by sequencing (ChIP-seq) technique has transformed the annotation of regulatory elements [135-137]. Through the ENCODE project [30, 138], there is a widespread access to millions of positions at which TFs are assumed to be present in at least one tissue or cell-type. The analysis of ChIP-seq data for the purpose of regulatory variant discovery has been introduced. In short, by combining large-scale genotype data (such as whole genome sequencing, WGS) with ChIP-seq, it is now feasible to identify the TF binding preference between the two alleles at heterozygous sites within bound regions [49, 50, 132, 139, 140]. Heterozygous site binding events can be classified as allele specific binding (ASB) or non-ASB events specifying whether one allele is significantly preferred or not. The advantages of heterozygous site binding data are that: 1) it provides high-throughput compilation of altered TF binding data; and 2) it

compares TF binding at two similar sequences (one nucleotide difference) in the same cell context, reducing technical and biological noise [47].

In this work, we focused on heterozygous site binding events to interpret the impact of variations on TF-DNA binding. Using genotype calls from WGS, we extracted heterozygous site binding events across 45 TF ChIP-seq experiments from the ENCODE project [30]. We identified a set of features correlated with TF heterozygous site binding events, including motif alterations of the ChIP'ed TFs and other potential partner TFs, allelic difference of DNase I hypersensitivity (DHS), and allelic difference of histone modifications. Finally, a classifier was trained to predict the variation impact on TF binding, revealing that combining DHS and WGS was an efficient approach to predict altered TF binding. Our results suggest that heterozygous site binding events provide a foundation to identify features that informed the detection of *cis*-regulatory variants.

2.2 Materials and methods

2.2.1 Genotype data of investigated cell lines

Genotype data for GM12878 and six other lymphoblastoid cell lines (Table 2.1) were obtained from the Complete Genomics website (as of June 2014, specific hyperlinks provided in Appendix Table A1) [141]. For HeLa-S3, NIH granted permission to access raw sequence data (accession number phs000640.v2.p1) [142]. Encrypted SRA files of HeLa-S3 were converted to raw reads using fastq-dump command from sratoolkit (https://github.com/ncbi/sratoolkit, Version 2.3.2). Raw reads were mapped to the hg19 reference genome using bwa (version 0.7.10-r789) with the command bwa sampe with default parameters. The GATK tool (version 2.7-4-g6f46d11) IndelRealigner [143] was used to realign reads around indels. Finally, samtools

(version 0.1.9-r783) mpileup [144] was used to call variations. Any variation with a quality of at least 30 was kept for subsequent analysis.

2.2.2 ChIP-seq read alignment

We downloaded ChIP-seq data for diverse TFs, DHS data, and histone modification data, from the ENCODE project [30] (Appendix Table A1). For each cell line, we built a personalized version of the hg19 reference genome in which the single nucleotide variation (SNV) sites were replaced with IUPAC degeneracy codes according to the genotype data. The downloaded ChIP-seq reads were mapped to the personalized reference genome using Novoalign (version 3.01.00) with default parameters. We removed any reads with a mapping quality lower than 30.

2.2.3 Mapping bias simulation

Even though we used a personalized reference genome to improve the mapping sensitivity of alternative alleles, there remained a potential mapping bias towards certain alleles (e.g. reads with strong similarity to multiple genome regions) [49, 53, 132]. To address this issue, we performed a read mapping simulation to estimate the mapping bias at each heterozygous site. For each heterozygous site within the TF ChIP-seq peak regions, we generated all the possible 36-bp reads overlapping with the heterozygous sites for each allele and each strand. Then, the generated reads were mapped to the personalized reference genome using the same settings as for the real ChIP-seq data. Finally, we assessed the mapping bias and excluded the biased sites if the imbalance ratio of any allele was greater than 60%. This filter threshold is set according to the minimum imbalance ratio threshold for called ASB events (see below).

2.2.4 Retrieving heterozygous site binding events and calling ASB events

Uniformly processed ChIP-seq narrowPeaks were downloaded from ENCODE [145]. In order to increase the confidence of TF-bound regions, we narrowed the peaks to the 100 base pairs (bp) core regions centered around the peak max positions [146, 147]. For each TF ChIP-seq data set, we retrieved the read counts of the two alleles at heterozygous site binding events within the ChIP-seq peak core regions. Replicates were pooled together to increase the overall read coverage [148]. Peak core regions on sex chromosomes or overlapping with copy number variant regions (Appendix Table A1) were filtered out. We excluded from the downstream analyses core peaks that harboured multiple heterozygous SNVs (8,311 out of the 79,565 heterozygous core peaks) to ensure that the two alleles differed by a single nucleotide. For the sites supported by at least 10 reads, an ASB event was called if the read count on one allele was significantly different from the other allele based on a binomial test (false discovery rate, or FDR, <0.05). As an aside, we explored the option of using replicate normalization (see Appendix A1and Appendix A2). The hypothesized probability of the binomial test was set as the mapping imbalance detected in the above-mentioned read mapping simulation at each heterozygous site. For ASB events, we further required the favored allele to show at least 60% allele imbalance (proportion of reads mapped to one allele over the total) following [148], to remove extreme p-values caused by small changes at high read depth *loci*. We labelled the allele with higher number of mapped reads as the favored allele, and the lower one as the unfavored allele; in non-ASB events, if the numbers of mapped reads on the two alleles were equal, the reference allele was labelled as favored.

2.2.5 TFBS identification in ChIP-seq peak regions

TF binding motifs were downloaded from the JASPAR database (version 2014) [149]. The motif of each corresponding TF was scanned against the peak regions using the Biopython (version 1.65) motifs module [149, 150]. For each scanned site, the motifs module provided the motif score (position weight matrix, PWM, score) and the p-value of the score against a null uniform distribution of the four nucleotides (referred to as motif p-value). Sites with scores above the false positive rate of 0.001 were predicted as TFBSs.

2.2.6 Defining ASB frequency within TFBSs

We defined the frequency of ASB events at each TF motif position as the proportion of ASB events observed at this position over the total number of ASB events observed across all motif positions considering the predicted TFBSs; the same definition was applied to non-ASB events. Only the TFs with at least 10 ASB and 10 non-ASB events in the predicted TFBS are considered to calculate the ASB frequency within TFBS.

2.2.7 Identifying comotifs within ChIP-seq peak regions

We used the findMotifsGenome.pl script from the HOMER [26] package (version 4.6) with default settings to identify enriched known motifs in ChIP-seq peak regions. The HOMER default analysis window of 200bp was applied. Among the enriched motifs reported by HOMER, we identified the five most enriched motifs according to the following criteria: 1) not similar to the motif of the ChIP'ed TF if available in JASPAR; and 2) no similar motifs within the five identified motifs. Motif similarity was based on the compare-matrices command provided in the RSAT toolset (version 2011) [151] with an information content correlation threshold of 0.8. For

the ASB SNVs not overlapping the predicted TFBSs of the ChIP'ed TF, we tested the correlation between motif alteration (log ratio of motif p-values between the two alleles as in [70]) and allele imbalance of TF binding within the predicted TFBSs of each of the five enriched motifs (Spearman correlation, FDR < 0.05). The significantly correlated enriched motifs were identified as comotifs.

2.2.8 Association between cobound TFs and ASB events

To identify the distribution of ASB events within binding regions of other TFs, we used all the available TF ChIP-seq peaks in the same cell line. Cobound TFs were identified if their peaks overlapped with the peaks of ASB TFs. For the heterozygous site binding events of each ASB TF, we investigated the association between the presence of ASB events and their overlap with the peaks of each cobound TF (two-sided Fisher's exact test, FDR <0.05). Throughout the thesis, the minimum reported p-value is 2.2×10^{-16} which is based on floating point constraints. The odds ratio of Fisher's exact test was used to interpret whether ASB events were enriched (odds ratio >1) or depleted (odds ratio <1) in cobound regions.

2.2.9 Classification of heterozygous site binding events

We used the randomForest package [152] and the recursive feature elimination function from the caret package [153] to train random forest classifiers ("ntree" parameter was set to 1000) and select key features. Since there were more non-ASB events than ASB events, non-ASB events were randomly downsampled to balance the training data set for each tree building process following the balanced random forest approach [154, 155]. We used a five-fold cross-validation approach to assess the predictive power of the classifiers. Specifically, the predictive power

corresponded to the average area under precision-recall curve (AUPRC) obtained through the five-fold cross-validation. For determining the importance of each feature in a classifier, we took the "MeanDecreaseAccuracy" (mean accuracy decrease over all trees) score reported by the random forest.

The input features, listed in Appendix Table A5, spanned five categories: 1) motif-related features, for instance the motif scores of the two alleles, the best motif scores within the peak regions on two alleles; 2) positional information, such as SNV distance to the ChIP-seq peak max and SNV position within the predicted TFBS; 3) enriched-motif related features (log ratio of motif p-values between the two alleles); 4) cobound TFs, such as the overlapping of heterozygous site binding events with each available cobound TF peaks within the same cell line; and 5) chromatin features, for instance the read counts on the two alleles from DHS and 11 histone modification data from the corresponding cell type. We combined features across the five categories and trained three models: 1) a Seq model based on sequence features, including categories 1-3; 2) a Seq+DHS model adding DHS data on top of the Seq model; and 3) a Full model trained using all features.

We compared our classifiers to deltaSVM [40] and BayesPI-BAR [73]. The deltaSVM score was calculated as the gkmSVM score difference between two alleles [40]. For each TF, we trained a separate gkmSVM model (version 2.0) with default parameters using 5,000 randomly selected ChIP-seq peaks following [37] and the associated tutorial (http://www.beerlab.org/gkmsvm/). One TF (PRDM1) had only 4,577 peaks and we used all of them to train the gkmSVM model.

The BayesPI-BAR package was downloaded from http://folk.uio.no/junbaiw/BayesPI-BAR/, and BayesPI-BAR scores were calculated with default parameters.

2.3 Results

2.3.1 Compile heterozygous site binding events

We implemented a pipeline that combined ChIP-seq and genotype data from the same cell types to extract heterozygous site binding events (Materials and methods in Chapter2). Specifically, ChIP-seq (and DHS) reads were mapped to personalized reference genomes in which the variants reported in the genotype data were incorporated. In total, we retrieved 51,518 heterozygous site binding events supported by at least 10 reads from 45 TF ChIP-seq data sets from eight cell lines. We also extracted read counts of 11 histone modifications and DHS on the two alleles of TF heterozygous site binding events in GM12878 and HeLa-S3 cell lines. We observed that 4.3% of the TF ChIP-seq peak regions contained a single heterozygous site (Table 2.1 and Appendix data). ASB events were defined if the number of mapped TF ChIP-seq reads on one allele was significantly higher than the number of mapped reads on the other allele (Binomial test, FDR < 0.05) and with at least 60% allele imbalance for the favored allele as in [148]. We found that 20.9% of heterozygous site binding events were classified as ASB events; others were classified as non-ASB events. Among the compiled data of eight cell lines, GM12878 and HeLa-S3 (Tier 1 and Tier 2 cell lines from the ENCODE project) had data sets for all the investigated TFs, DHS, and histone marks; the remaining six cell lines were restricted to ChIP-seq data for CTCF. Therefore, we focused on GM12878 and HeLa-S3 for most of the study, using the additional cell lines for testing the classification models (see below).

Cell	TF	DHS and Histones	Peak Count	Heterozygous binding sites events	ASB
GM12878	16	12	405,427	17,222	2,314
HeLa-S3	23	12	518,558	18,481	5,533
GM12872	1	0	47,151	2,496	488
GM12873	1	0	51,005	2,575	552
GM19238	1	0	49,938	2,909	500
GM19239	1	0	41,085	2,473	282
GM19240	1	0	46,036	2,972	573
GM12864	1	0	46,798	2,390	523
Total	45	24	1,205,998	51,518	10,765

Table 2.1 Overview of heterozygous site binding data

For each investigated cell line (first column), we report the number of compiled TF ChIP-seq experiments (second column) and DHS and histone modification data sets (third column). The corresponding total number of TF ChIP-seq peaks is given in the fourth column. Finally, we provide the number of heterozygous sites supported by at least 10 reads within the ChIP-seq peaks (fifth column) and the number of ASB events (sixth column). Note that the numbers are derived from the compilation of all the TF ChIP-seq data for each cell line. Details for each TF can be found in Appendix Table A2.

2.3.2 TFBS alterations strongly correlate with ASB events

To understand the underlying genetic mechanisms of ASB, we considered the subset of SNVs overlapping with the predicted TFBSs (Materials and methods in Chapter2). An initial analysis revealed that ASB SNVs were significantly enriched in predicted TFBSs compared with non-ASB events (p-value $< 2.2 \times 10^{-16}$, odds ratio = 3.0, Fisher's exact test). Next, we assessed the motif score alteration caused by the SNVs for ASB events. We found that motif scores of

favored alleles (allele with higher read count) were significantly higher than those of unfavored alleles in predicted TFBSs (Figure 2.1, p-value < 2.2×10^{-16}, estimated median difference = 0.04, one-sided Wilcoxon signed-rank test), reflecting the contribution of motif score alteration to ASB events. In contrast, non-ASB events displayed a balanced score distribution between the two alleles. Our results agree with previous observations [47, 48] but are based on data for a much larger number of TFs and TFBSs. However, only a portion of ASB SNVs (19.3%) overlapped with the predicted TFBSs, indicating that additional mechanisms beyond TFBS alteration contribute to the observed ASB events. A plot showing the total set of ASB and non-ASB events, including those outside the predicted TFBSs is provided in Appendix Figure A1.

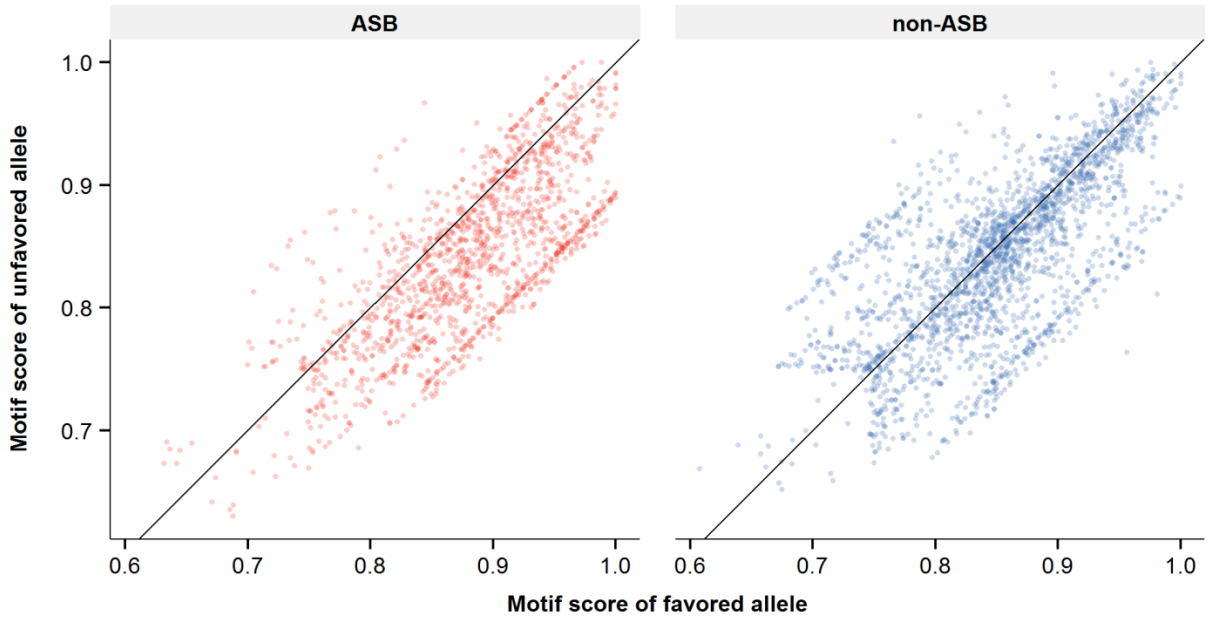


Figure 2.1 TFBS motif score analysis at heterozygous site binding events

In each panel, we plot the motif score at heterozygous sites on the favored allele (harbouring higher amount of mapped ChIP-seq reads, x-axis) and unfavored allele (y-axis) at predicted TFBSs. ASB (left panel) and non-ASB

(right panel) events are plotted separately. The black diagonal lines indicate an identical motif score on the two alleles. Note that the figure was generated using all heterozygous site binding events for all compiled TFs in GM12878 and HeLa-S3.

2.3.3 ASB events show different positional distribution within TFBS compared with motif information content

We next examined whether specific positions within TF binding motifs were more sensitive to ASB events and how such impactful positions related to their information content (IC) in the TFBS motif profiles. IC has been correlated with the strength of binding site preference for individual nucleotides in TF binding models, and the maximum IC of a position is two bits when a certain nucleotide is consistently observed at that position in TFBSs [17]. Given a TFBS motif, the positional impact was measured as the frequency difference between ASB and non-ASB events at each position (Materials and methods in Chapter2). As expected, positional impact was significantly correlated with positional IC across motif positions of all investigated TFs (Spearman correlation coefficient = 0.38, p-value = 6.6×10^{-12}; Figure 2.2A). But most motif positions did not strictly follow this trend in Figure 2.2A, revealing a large variance of positional impact that cannot be attributed to IC.

The most extreme cases at the upper right corner of Figure 2.2A represented motif positions where TF binding was disproportionately impacted. We qualitatively observed that these positions tended to be centrally positioned within the TFBSs of the TFs which were dimers and bound symmetrically to DNA. When analyzing all four symmetric TF dimers in our data sets with known TF-DNA complex structures (CEBPB, MAX, TCF7L2, and USF1), we observed

that central positions significantly showed high positional impact compared with other positions with similar IC (p-value = 0.02, estimated median difference = 0.09, one-sided Wilcoxon rank-sum test). As a specific example, CEBPB recognizes an 11bp motif containing four positions with an IC of two bits (positions 3, 4, 6, and 10), which, according to the motif, would be expected to be equally important for binding (Figure 2.2B). However, the positional impact was particularly high at position 6, at the center of the motif, indicating that this position could be more critical for the disruption of TF binding (Figure 2.2B). Further structural analysis of a DNA-CEBPB dimer interaction revealed that position 6 was contacted by both monomers (Figure 2.2B). The critical role of central positions suggests that mutations at these positions might potentially affect the binding of the two monomers. Recently, the same position of the CEBPD motif was reported to display more somatic mutations within the predicted TFBSs than other positions in human cancer genomes [156], which is concordant with our findings. Other cases included the PAX5 motif at position 15 (Appendix Figure A2), which was of low IC (0.4) but with high impact, suggesting that low IC positions could also be critical for TF binding [157]. Taken together, IC derived from motifs partially explained the distribution of ASB events across the motif, while positional impact from ASB events provided deeper insights into the binding properties of TFs.

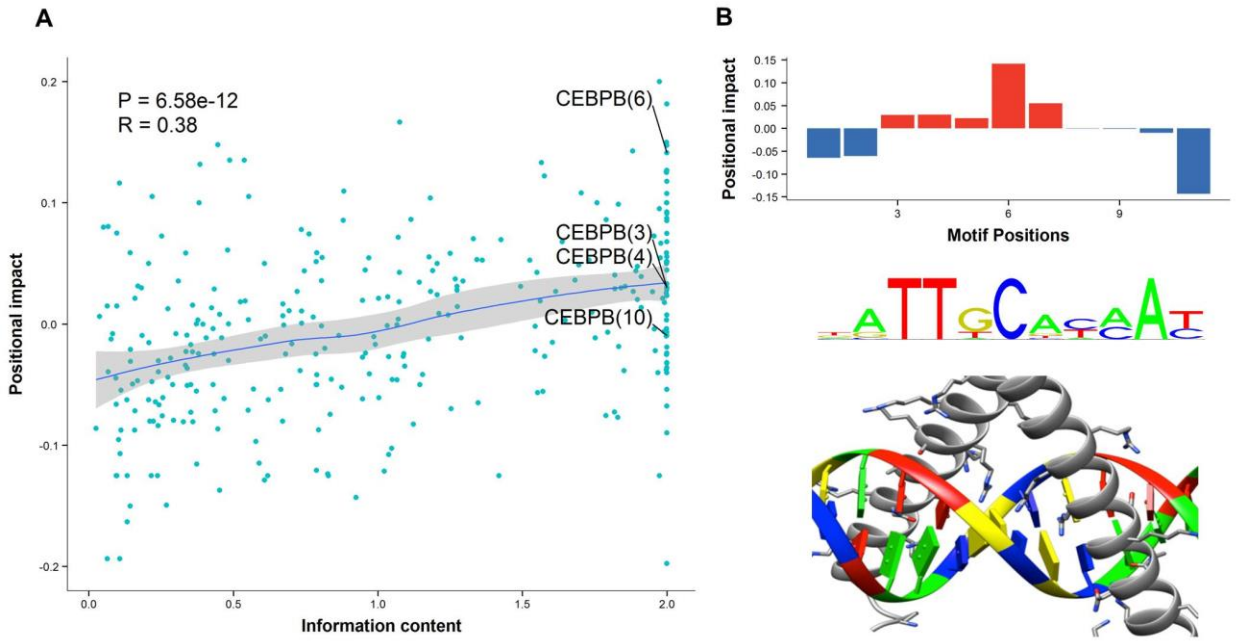


Figure 2.2 Information content and positional impact of each position within TFBS

(A) Correlation between positional impact and information content. Each point corresponds to a position within TFBSs associated to ChIP'ed TFs, and four positions are explicitly labeled in parenthesis for CEBPB. Positions are plotted with respect to their associated information content (x-axis) from the TF motif and positional impact (y-axis). The trend line is drawn by the locally weighted scatterplot smoothing method. (B) Exceptional example of CEBPB motif with its positional impact distribution (upper), TF binding motif logo (middle), and TF-DNA interface (lower; Protein Data Bank ID: 2e42).

2.3.4 Disruption of enriched comotifs can lead to ASB events

Since most variations at ASB events were outside of the predicted TFBSs (80.7%), we assessed whether disrupted TFBSs of potential partner TFs could be responsible for the observed events. We retrieved the five most enriched, non-redundant motifs within the peak regions of each TF ChIP-seq experiment (Materials and methods in Chapter2). Within the predicted TFBS of each enriched motif, we tested the correlation between the motif score change and the allelic binding

imbalance of the ChIP'ed TF across all heterozygous site binding events (Materials and methods in Chapter2). We found fifteen significantly correlated enriched motifs for nine TF ChIP-seq experiments (based on the Spearman rank statistic, FDR < 0.05, Figure 2.3), hereafter referred to as comotifs. Decreased motif scores of comotifs were preferentially observed on unfavored alleles in ASB events, consistent with a cooperative binding model [56]. The comotifs lay in three categories (Appendix Table A3): 1) seven cases in which the TFs associated to the comotifs were known to interact with the ChIP'ed TF, for instance the comotif of P300 was CEBPB (P300-CEBPB); 2) one case (RUNX3-RUNX1) in which the TF of comotif belonged to the same TF family as the ChIP'ed TF; and 3) seven cases of novel relationships, from our knowledge, including CEBPB-BATF, and P300-NF-E2.

Moreover, six out of the fifteen comotifs arose from the experiments in which the ChIP'ed TFs did not bind DNA directly, for example P300. For these non-sequence specific TFs, 33.5% of ASB-SNVs overlapped the TFBSs of comotifs, significantly enriched compared with 17.4% for non-ASB events (p-value < 2.2×10^{-16}, odds ratio = 2.4, Fisher's exact test, Appendix Figure A3). Overall, ASB overlapping comotifs comprised 9.4% of ASB events.

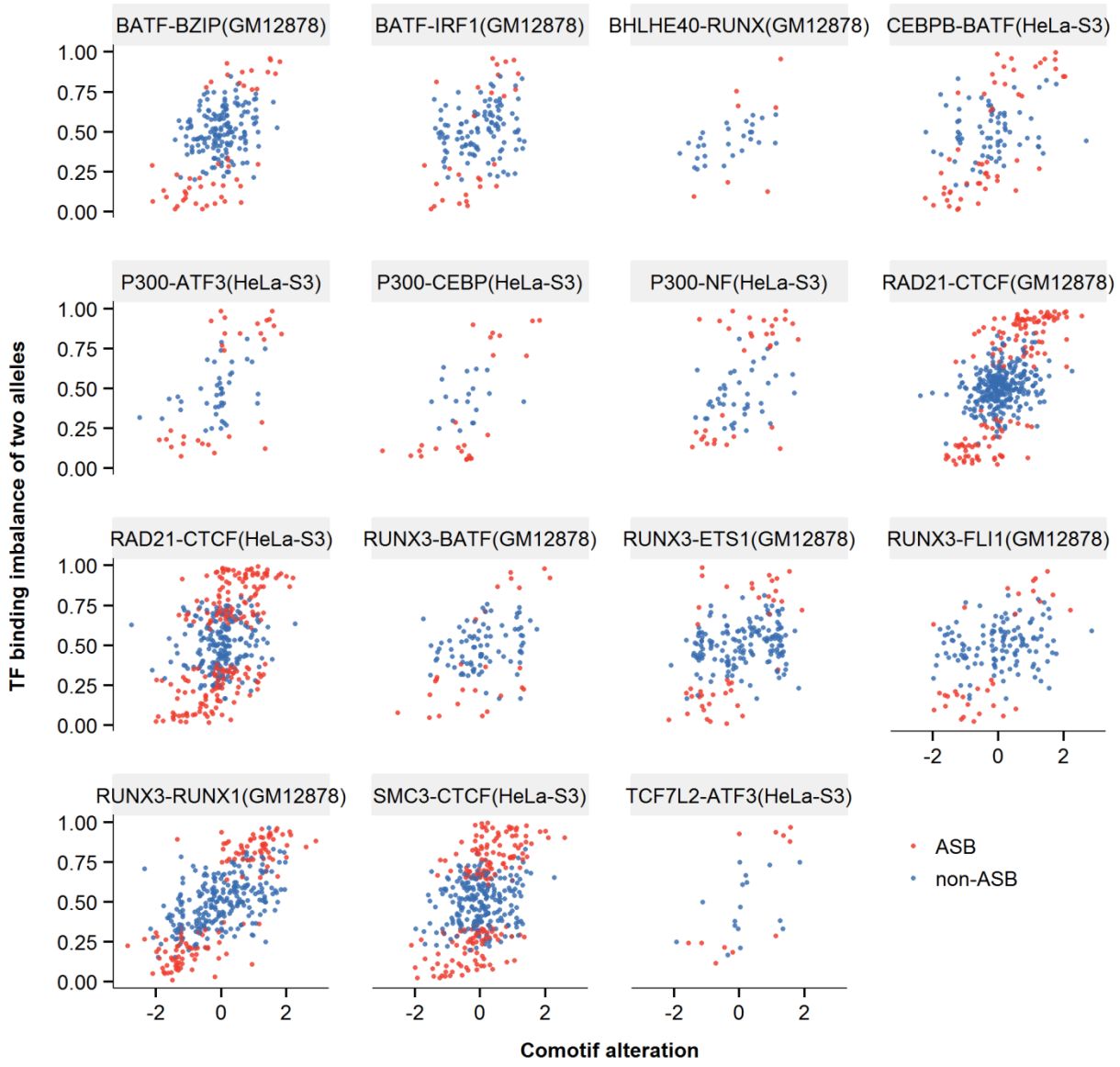


Figure 2.3 Alteration of comotif correlated with TF allelic imbalance

The name of each panel specifies the ChIP'ed TF followed by the comotif name and the cell line in parentheses. Each dot represents one heterozygous site binding event (red for ASB and blue for non-ASB events) found within the predicted TFBSs of the comotif. The comotif alteration (x-axis) represents the log ratio of motif p-values between the reference and alternative alleles. The allelic binding imbalance (y-axis) indicates the fraction of reads mapped on the reference allele over the whole read coverage at that position. We test the correlation between the

two properties for each ChIP'ed TF and its enriched HOMER motifs, and only significantly correlated pairs are plotted (FDR < 0.05).

2.3.5 ASB events are associated with cobound TFs

Next we sought to understand how ASB events related to regions bound by additional TFs within the same cell using ChIP-seq data. It has been observed that TF binding in cobound regions (cases where ChIP-seq data for multiple distinct proteins have overlapping peaks) tends to be more conserved over evolution than isolated binding events [158]. We tested the distribution difference between ASB and non-ASB events in the ChIP-seq peaks of each cobound TF (Materials and methods in Chapter2), revealing 106 significant pairs (Appendix Table A4, Fisher's exact test, FDR < 0.05). Of these, 47 were observed in GM12878 lymphoblastoid cells, and almost all (46 out of 47) displayed depletion of ASB (relative to non-ASB) in the cobound regions (odds ratio <1). This pattern is concordant with the concept of variant buffering effects in motif-rich DHS regions [148]. For instance, CTCF heterozygous site binding events were classified as ASB in 8.9% of cases where ZNF143 binding peaks were overlapping, while 18.3% of cases were classified as ASB if there were no overlapping ZNF143 peaks (p-value $=3.6\times10^{-11}$, odds ratio = 0.43, Fisher's exact test). The ASB TF and cobound TF pairs included known TF-TF interactions, such as CTCF-ZNF143, and RUNX3-YY1 [159], suggesting functional interactions for the pairs observed. In HeLa-S3, a cancer cell line, we observed a reversed pattern where ASB events were enriched in cobound regions (odds ratio >1, not depleted as in GM12878) for 35 out of 59 cases (such as CEBPB-P300 and MAX-CMYC). The opposing pattern between normal and cancer cells suggests that binding site alterations in cobound regions

of cancer cells may be functionally important for gene dysregulation. Further analyses would be required to test this hypothesis when more TF binding data become available.

2.3.6 Allelic chromatin properties coordinate with ASB events

To further shed light to the mechanisms associated with ASB events, we investigated the non-genetic properties in proximity to ASB events. We extracted read counts from DHS and histone modification ChIP-seq experiments on the two alleles at heterozygous site binding events. Next, we assessed the correlation between allelic imbalance of each chromatin property (DHS and 11 histone modifications) and TF binding. Overall, 196 significant correlations were observed (Pearson correlation, FDR < 0.05; Figure 2.4 and Appendix Figure A4). DHS signal was significantly correlated with TF binding for 35 out of 39 TF ChIP-seq experiments. DHS showed higher read counts on the TF favored allele for 73.4% of the ASB events compared with 52.5% for non-ASB events. Moreover, we found 161 TF-histone correlation pairs. Active histone modifications, such as H3K27ac, H3K4me2, and H3K3me3, exhibited positive correlation patterns with TF binding imbalance. Taken together, DHS and histone modifications widely correlated with ASB events, indicating their potential value for predictive modelling.

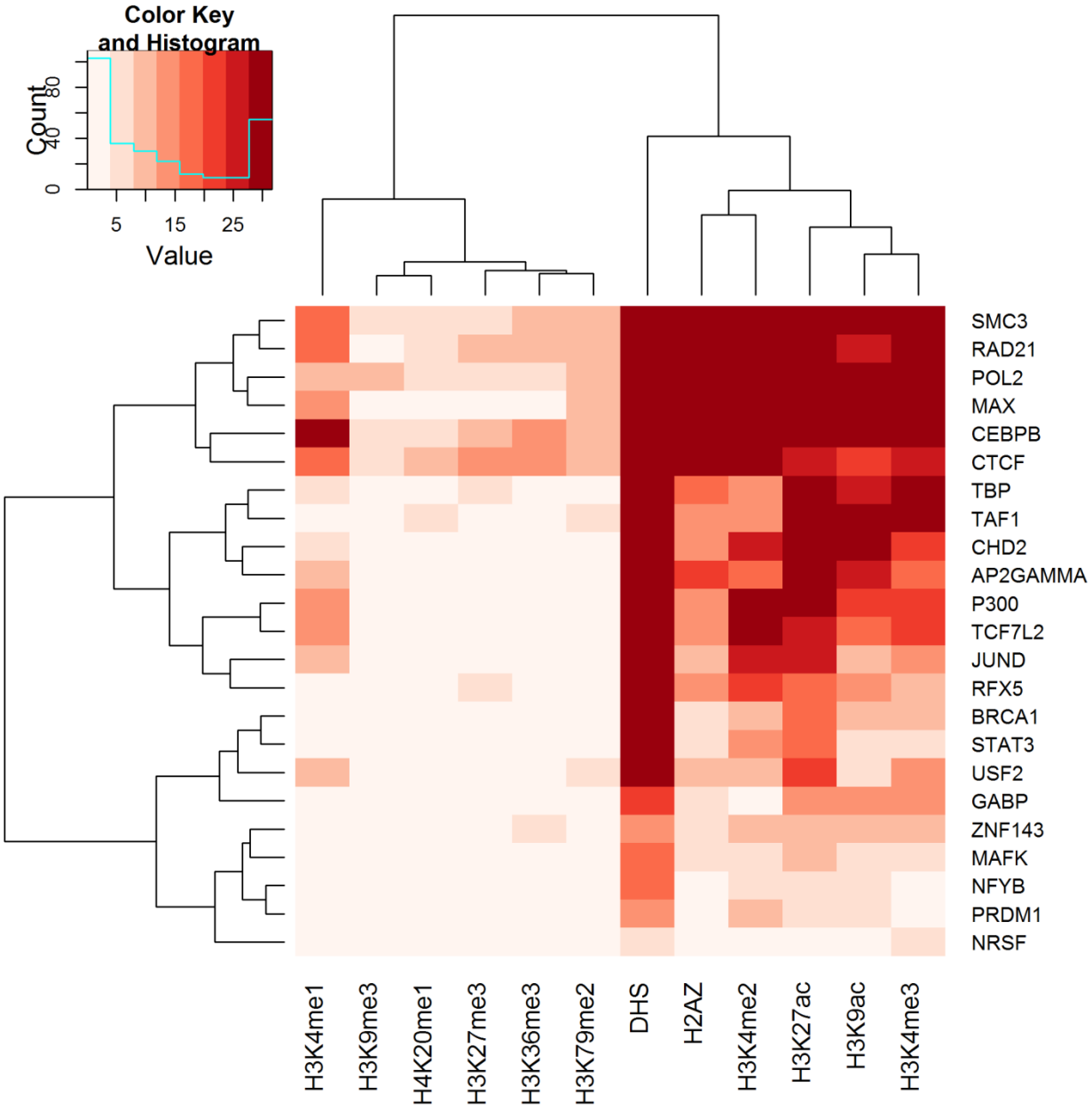


Figure 2.4 Allelic coordination between TFs and chromatin properties in HeLa-S3

The heatmap represents the -log(p-value) of Pearson correlation between allele imbalance of TF ChIP-seq reads at heterozygous site binding events and chromatin properties (DHS and histone modifications).

2.3.7 DHS and sequence-derived properties are sufficient for cost-effective ASB event prediction

Building upon the observed associations between ASB events and properties of both sequence and experimental data, we constructed computational models to determine our capacity to predict SNVs disruptive of TF binding (that is to distinguish between ASB events and non-ASB events). We took ASB events as the positive training data and non-ASB events as the negative set for model training. We constructed random forest classifiers using only DNA sequence information (that are the features derived from motif and comotifs, referred to as Seq model, see Materials and methods in Chapter2) and assessed their predictive performances. Consistent with past literature [69, 70, 73], the Seq model had predictive value but the performance was quite limited across all the investigated TFs (average AUPRC of 0.35, Figure 2.5A). The Seq models allowed consistent performance across data from multiple individuals within the same cell type (Appendix A3).

We compared our classifiers against two existing sequence-based models, deltaSVM [40] and BayesPI-BAR [73]. Seq models outperformed BayesPI-BAR (p-value = 7.5×10^{-9}, estimated median AUPRC difference = 0.18, Wilcoxon signed-rank test, Appendix Figure A6) and showed similar performance with deltaSVM (p-value = 0.33, Wilcoxon signed-rank test, Figure 2.5A) when predicting ASB events. The differences between deltaSVM and our Seq models are that deltaSVM uses *k*-mers to predict TF-bound regions while our Seq models allow for combining positional feature on top of motif features (Materials and methods in Chapter2). The ASB framework potentially can incorporate any features of two alleles into the discriminative model, for example, adding deltaSVM scores to cover the *k*-mer changes.

Next, we took into account all the features analyzed in the previous sections into the model (Materials and methods in Chapter2), which was hereafter referred to as the Full model. The Full model showed a mean AUPRC of 0.43 across all the tested TFs (Figure 2.5A). For those TFs with known binding motifs, the top ranked features highlighted two major categories contributing to the success of the model, DHS and motif sequence properties. Specifically, the top five features were DHS read count from the unfavored allele, DHS read count from the favored allele, motif score on the unfavored allele, motif score on the favored allele, and H3K27ac read count from the favored allele (Figure 2.5B). For TFs lacking a motif model, the feature set could not include motif sequence properties of the ChIP'ed TF. Consequently DHS, H3K4me2, and H3K27ac were important for the success of the classifiers (Appendix Figure A7).

Given that ChIP-seq TF binding data were not available for most cell lines, while DHS was more likely to be available, we evaluated the performance of models limited to sequence-derived features and DHS (Seq+DHS model). Consistent with the number of features in the training sets for each model, results showed that the Full model outperformed the Seq+DHS model, which in turn outperformed the Seq model across all the tested TFs (p-values of 2.1×10^{-8} and 1.3×10^{-5}, Wilcoxon signed-rank test) (Figure 2.5A). From a sequence-only baseline of 0.35 in terms of average AUPRC, the Seq+DHS model achieved 0.40, and the Full model achieved 0.43. Importantly, inclusion of DHS with sequence properties provided important value, representing 62.3% of the average improvement of the Full model over the sequence-only baseline. These results highlighted that ASB prediction could be pursued with few laboratory generated features cost-effectively by coupling sequence analysis with experimental genotyping (WGS) and DHS data.

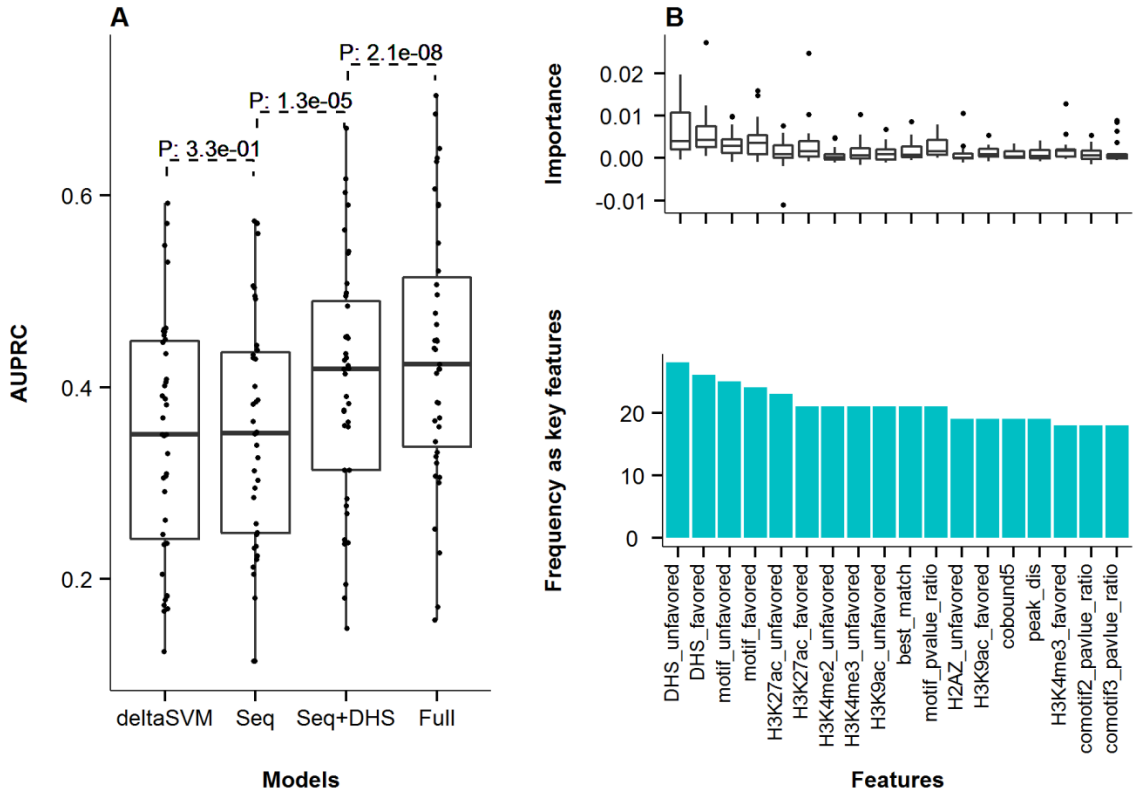


Figure 2.5 Performance of ASB classification models and key features

(A) AUPRC of the deltaSVM, Seq, Seq+DHS and Full models across the 39 investigated TF ChIP-seq experiments. Seq models are based only on sequence-related features; Seq+DHS models include DHS data on top of the Seq model; and Full models further include histone marks and cobound TFs. Each dot represents the model performance of one TF ChIP-seq experiment in one model. Details on each model and features can be found in Materials and methods in Chapter2. (B) Top frequent key features in the Full models for all 27 TFs with known motifs. The suffix 'favor' and 'unfavor' refer to the favored and unfavored alleles at heterozygous sites. The 'motif_pvalue_ratio' is the log ratio between two alleles in terms of motif score p-value. The 'peak_dis' indicates the distance of the SNV to ChIP-seq peak max position where the highest number of reads are mapped within the peak.

2.4 Discussion

Predicting variant impact on TF binding is amongst the biggest current challenges for genome interpretation. One of the main obstacles is the lack of sufficient and reliable TFBS alteration data, which are critical for the development of bioinformatics methods. We compiled 10,765 ASB events from 45 TF ChIP-seq experiments from eight cells lines. To the best of our knowledge, this is the largest experimentally defined ASB collection. While altered canonical TFBSs for the ChIP'd TFs were frequently observed (19.3%), most ASB SNVs did not overlap with the primary TF motif. When looking across positions within TFBS, we observed that central TFBS positions for symmetric TF dimers were more critical than other positions with similar information content. Alterations of comotifs, potentially bound by partner TFs, were observed for a portion of ASB events (9.4%). Taking the enlarged collection of data to train classification models, we demonstrated that baseline models using only genomic sequence data were improved by the incorporation of allelic DHS data, which provided 62.3% of the performance improvement achieved by models using all available features (~100 per cell type) from the ENCODE data.

The applied thresholds in ASB calling were based on previous studies and justified in the context of our data. As we were interested in both ASB and non-ASB events, a moderate threshold (FDR < 0.05) was applied to strike a balance between the two types of events as in [47 , 148, 160]. To avoid heavy test burden, ASB studies also filtered low coverage sites that lack statistical power in the binomial test. However, the threshold for minimum read coverage differs across studies (e.g. 7 reads in [47], and 10-20 reads adjusted by library size in [48]). Based on our datasets, we

identified a threshold of 10 reads, which is the minimum read coverage for the most imbalanced heterozygous sites (all 10 reads on one allele) to reach ASB significance threshold.

Our results suggest that positions of SNVs within TFBSs should be considered when investigating SNV impact on symmetric TF dimers. The observed impact of SNVs within these central positions was not fully reflected by the information content of classic motif (position weight matrix) models [31]. Classic PWM-based methods [69, 70, 72, 73] did not capture such characteristics when predicting TF binding alteration. The importance of these central positions was supported by structures of DNA-TF dimer interactions showing them to be dual-contact points for both protein subunits, highlighting that structural information can be important for understanding the impact of SNVs on TF binding.

Our ASB classification model provides a novel supervised and integrative framework to model SNV impact on TF binding. To evaluate the impact of SNVs, most prior methods calculated binding score differences between altered alleles and reference alleles based on TF binding motifs [70, 72, 73] or enriched *k*-mers [38, 40]. Prediction of SNV impact was based on those cases where the difference exceeded a threshold. However, the selection of a threshold was difficult to justify. In contrast, our ASB model learned the optimal threshold (decision surface) from the data directly. Moreover, our method was not limited to sequence features (TF motifs and *k*-mers), with the capacity to incorporate diverse features (such as genetic features, DHS, and histone modifications). We anticipate that such features will become increasingly available in the near future. In addition, the relative importance of each feature in the classification models provided insights into the mechanisms contributing to TF binding.

Only ~30% of ASB events can be explained by motif or comotif alteration. Understanding how the altered binding arises in the remaining portion is likely to require advances in our knowledge and understanding of TF binding. First, the available TF binding models are insufficient. Most human TFs do not yet have binding models, although the coverage improves [39]. Second, the existing binding models can be improved. For instance, CTCF has been shown to recognize flanking motifs that stabilize binding, but these are not yet well represented in the model [161]. Moreover, there are properties outside the sequence-specific target that contribute to binding. Flanking sequences can influence binding strength [162-164], potentially involving the shape (topology) of DNA [165, 166]. As we advance our understanding, we can anticipate that the causally unexplained portion of ASB events will be decreased. Overall, we recognize that there is an upper limit for DNA sequence to explain ASB events as other features can also contribute to TF binding, such as chromatin accessibility and DNA methylation.

The predictive power (AUPRC) of the ASB classification models is limited, particularly when considered on the scale of analyzing a full genome. The inadequate performance might be attributable to multiple causes. For instance, the classification model may be under-fitted because the number of ASB events available for training was not sufficient. Recently, two studies compiled new ASB datasets in other cell lines to investigate GWAS loci or the variant impact on gene expression [132, 167]. In the future, we anticipate a rapidly growing body of ASB data will be critical in training more reliable models. Alternatively, the set of features available for modeling may have missing components, e.g. the limited set of TF binding models. Lastly, ASB events could be caused by multiple SNVs or distal SNVs. In our data compilation, we excluded

the cases where multiple heterozygous SNVs situated within the same ChIP-seq core peak regions to simplify the analysis. However, the accumulated effect of multiple SNVs proximal or distal to a TFBS could alter local TF binding according to the TF-TF interaction and chromatin interaction models [128, 168]. Further efforts needs to be devoted to these areas.

Identification of *cis*-regulatory variants is a critical need for understanding the genetic mechanisms contributing to diseases [27]. Our compilation of heterozygous site binding data and ASB classification models provide unique data sets and a novel framework for modeling the impact of SNVs on TF-DNA interaction. Future advances in sequencing technology and enlarged ASB database will enable the reliable identification of *cis*-regulatory variants.

Chapter 3: Evaluating five statistical methods to call allele specific binding events

In Chapter 2, the assessment of candidate ASB events was performed using a statistic based on an assumption of a binomial distribution. As has been explored for determining the significance of RNA-seq differential expression [133, 169], there are a variety of approaches that could be considered. Recent studies suggest that the distribution of observed allelic reads might not conform to a binomial distribution, and it has been proposed that a beta-binomial distribution might be more appropriately used. In this chapter, using the ASB datasets from the previous chapter, diverse ASB calling methods are evaluated.

3.1 Introduction

Disease-associated variants identified using GWAS studies are enriched in regulatory regions which control the expression of genes [11, 121, 128], but functional roles of individual variants remain unclear. A potential mechanism for a subset of these regulatory region variants is through disruption of TF binding sites (and consequently the modulation of downstream gene expression) [129]. However, predicting which variants will disrupt TF binding is an ongoing challenge in bioinformatics. A promising approach is to investigate the TF binding difference between two alleles at heterozygous sites [47, 49]. These heterozygous site binding events can be classified as allele specific binding (ASB) or non-ASB events depending on a significance threshold. ASB events have become an increasingly important source to interpret regulatory variants, as they quantitatively measure the TF binding difference between two alleles within the same cellular context.

To date, various methods have been developed to call ASB events based on TF binding data (e.g. ChIP-seq) and genotyping data (e.g. WGS) [49, 140]. The first step is to extract TF binding signal (mapped ChIP-seq reads) of two alleles at known heterozygous sites. Allelic binding signal can suffer from mapping bias when the reads from one allele are not mapped as properly as the reads from the other allele (i.e. the reference allele) [49]. To address this challenge, personalized genomes containing the observed alleles from the genotyping data can be used to improve ChIP-seq read mapping [49, 132], followed by read simulation to detect any remaining mapping bias [48, 52]. After extracting the signal from the two alleles, statistical tests are applied to call significant ASB events.

For the last step, assessing significance, key questions remain to be resolved. First, which statistical distribution is the most appropriate for the observed pattern of allelic reads? In the past, the binomial distribution has been widely used, but recent studies have reported an "over-dispersion" problem because the observed variance of allelic read counts is larger than expected in a binomial distribution [132] (Figure 3.1A). To overcome this problem, the beta-binomial distribution has been introduced to model extra variance in binomial distribution [132]. However, a thorough comparison between performances of the two distributions has yet to be performed. The second question pertains to handling ChIP-seq biological replicates. The majority of current approaches pool replicates, but the impact of this approach and alternative approaches have not yet been investigated. Third, which benchmark data is most optimal for evaluation of alternative approaches, as no gold standard ASB dataset has been established.

In this chapter, we try to address these questions by comparing five different ASB calling methods using our previously compiled heterozygous site binding events from Chapter 2. The five investigated methods cover three different statistical distributions (binomial, beta-binomial and negative binomial), and three approaches for handling replicates (pooling, normalization and joint probability). Our results suggest that five methods differ mainly in their statistical stringency, but they all provide similar significance rankings for heterozygous site binding events. We benchmark five methods based on allelic imbalance of DHS signal, with the method applied in Chapter 2 performing best (an approach that uses a binomial distribution and pools replicates). We demonstrate that the over-dispersion problem could be due to mild TFBS alterations, supporting that the binomial distribution is appropriately used as a null distribution in ASB calling.

3.2 Materials and methods

3.2.1 Datasets for evaluating ASB calling methods

We previously compiled 39 datasets of heterozygous site binding events by combining ChIP-seq and genotyping data from GM12878 and HeLa-S3 cells [52]. Each heterozygous site binding event includes: 1) the number of mapped reads on each allele from each ChIP-seq replicate of the investigated TF; 2) the number of DHS reads on each allele derived from DNase-seq experiment in the same cells as the ChIP-seq experiment; and 3) the estimated mapping bias by read simulation. Moreover, each heterozygous binding event contains at least 10 mapped reads between the two alleles. In total, we included 34,287 heterozygous binding events to benchmark the five different ASB calling methods.

3.2.2 Hypothesis testing for ASB calling

Calling an ASB event is to test whether the TF significantly prefers to bind to one allele relative to the other. Under the null hypothesis, both alleles of a heterozygous site are equally bound by the TF. We define allelic imbalance as the ratio of mapped reads on one allele over the total of mapped reads at the heterozygous site. ASB calling can then be converted to hypothesis testing as follows:

$$H_0\colon\ p = p_0\ \ vesus\ H_1\colon\ p \neq p_0 \tag{1}$$

where p is the allelic imbalance of the TF binding at the heterozygous site. By default, p_0 is 0.5 in the case of balanced TF binding. p_0 can be adjusted according to the mapping bias estimated by read simulation.

Under the above hypothesis testing framework, we compared five different methods (Table 3.1) to call ASB events. The five methods differ in two aspects: 1) the statistical distribution to represent allelic read counts (binomial, beta-binomial or negative binomial); and 2) the approach for handling ChIP-seq replicates (pooling, normalization or modeling the observed data with joint probability).

3.2.3 Binomial distribution for allelic reads

ASB events have been traditionally called based on the binomial distribution [47, 52, 148]. For a heterozygous site with *k* reads on one allele and a total of *n* reads, the probability of observing such event under the null hypothesis (balanced binding) is:

$$P(k|n, p_0) = \binom{n}{k} {p_0}^k (1 - p_0)^{n-k} \tag{2}$$

An ASB event is called if the number of mapped reads on one allele significantly differs from the expected number under null hypothesis (FDR <0.05).

3.2.4 Beta-binomial distribution for allelic reads

Recent studies have revealed that allelic read counts do not strictly follow the binomial distribution, showing greater variance than expected by chance in a binomial distribution (Figure 3.1A) [132, 170]. Some studies [132, 134] have used the beta-binomial distribution to model the observed over-dispersion in the data. Compared with the binomial distribution, the beta-binomial distribution uses an extra parameter to model the degree of dispersion. Ideally, this parameter should be estimated from the null distribution. In the following process, we used non-ASB events called with FDR >0.05 (binomial test) to represent the null distribution.

Given the null distribution, the probability of observing *k* reads on one allele and *n* reads in total at a heterozygous site can be formulated as:

$$P(k|n,p,\gamma) = P_{BetaBinomial}(k|n,p/\gamma^2,(1-p)/\gamma^2) \quad (3)$$

where *p* is the allelic imbalance of TF binding towards one allele, and γ is the dispersion parameter. The beta-binomial distribution is as follows:

$$P_{BetaBinomial}(k|n,a,b) = \binom{n}{k}\frac{B(k+a,n-k+b)}{B(a,b)}$$

where *a* and *b* are called shape parameters, and *B* is the beta function:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

In equation (3), both shape parameters are set according to the mapping bias obtained from the read simulation. For each TF ChIP-seq experiment, the dispersion parameter γ is estimated by the maximum likelihood approach across the null distribution using the VGAM package [171].

3.2.5 Pooling replicates for ASB calling

Each ChIP-seq experiment in our datasets usually has two or three biological replicates. In Binom+Pool and Beta+Pool methods, ChIP-seq replicates were pooled together to increase read coverage, and p-values for ASB calling were obtained based on pooled reads by testing the null hypothesis according to equations (2) or (3).

3.2.6 Modeling replicates by joint probability

To consider replicates independently, we introduced the joint probability scheme and likelihood ratio test from the methods of calling allele specific expression [134, 170]. Assume a TF ChIP-seq experiment has R replicates, and that a heterozygous site has a total of n_r reads and k_r reads on one allele in replicate r. The joint probability of the observed data $D = (k_r, n_r)_{r=1}^{R}$ is:

$$P(D|p,\gamma) = \prod_{r=1}^{R} Pr(k_r|n_r, p, \gamma_r) \quad (4)$$

where the p is the allelic imbalance of TF binding at the heterozygous site. Pr can be either the binomial (equation (2)) or beta-binomial distribution (equation (3)). In the case of applying the beta-binomial distribution, the dispersion parameter γ_r is estimated from each replicate (it is omitted for the binomial distribution). Finally, the hypothesis testing can be formalized as a likelihood ratio test:

$$\begin{cases} \Lambda = \dfrac{P(D|\,\hat{p},\hat{\gamma})}{P(D|p_0,\hat{\gamma})} \\ 2log\,(\Lambda) \sim \chi_1^2 \end{cases}$$

where $\hat{\gamma}$ is the estimated dispersion parameter, and $\hat{p}$ is the maximum likelihood estimate at the heterozygous site. Significance can be obtained through the test statistic $2log(\Lambda)$, which is asymptotically distributed as a chi-square distribution with one degree of freedom.

3.2.7 Negative binomial distribution to call ASB events with replicates

ASB calling can be understood as a problem of identifying differentially expressed genes in two conditions. Here two conditions refer to the two alleles of heterozygous sites bound by TFs. We used edgeR [133], a popular package for differential expression analysis, for ASB calling. edgeR normalizes the number of reads in each replicate and then uses the negative binomial distribution to identify differential signal between two conditions.

3.2.8 Using allelic imbalance of DHS signal to evaluate ASB calling methods

We have shown that allelic DHS signal widely correlates with allelic imbalance of TF binding at heterozygous sites [52]. Here, we evaluate the performance of five ASB calling methods based on their correlation with the allelic imbalance of DHS signal. Allelic imbalance of DHS signal was calculated as the ratio between the TF-favored allele (the allele with more mapped reads than the other in TF binding data) and total read depth. For each ASB calling method, we calculated the correlation between the allelic imbalance of DHS signal and the p-values resulting from ASB calling at the same heterozygous site for every TF data set. We focused on the heterozygous sites with certain read coverage (10 < read coverage < 20; Appendix Figure B1) as sites with higher read coverage tended to generate extremer p-values than the sites with the same degree of TF binding imbalance. Coefficients of significant correlations (FDR<0.05) were used to compare ASB calling approaches (Figure 3.2).

3.2.9 Scoring DNA sequence using PWM

Position weight matrices (PWMs) of TFs were derived from TF binding profiles in the JASPAR database (version 2014) [149]. The PWM of each corresponding TF was scanned against the candidate sequences using the Biopython (version 1.65) motifs module [149, 150]. For each scanned site, the motifs module provided the PWM score and sites with scores above the false positive rate threshold of 0.001 were predicted as TFBSs.

3.2.10 Code and data availability

ASB calling methods in this work are implemented in R [172]. The code and data links used in this work can be found at www.github.com/wqshi/asb_call.

3.3 Results

3.3.1 ASB calling methods provide highly correlated p-values but differ in statistical stringency

We implemented five different methods (Table 3.1) for calling ASB events and compared their called ASB events in 39 TF ChIP-seq experiments (see Materials and methods in Chapter3). We included a total of 34,287 events supported by at least 10 reads. Four of the evaluated ASB calling methods (Table 3.1) are combinations of two statistical models (binomial or beta-binomial) and two approaches to handle replicates (pooling and joint probability). The last approach, edgeR [133], is a popular method to identify differential signal in gene expression or TF binding analyses.

Approach	Statistical distribution	Replicates
Binom+Pool (traditional approach used in [52, 148])	Binomial	Pool replicates together
Beta+Pool [132]	Beta-binomial	Pool replicates together
Binom+Rep	Binomial	Calculate the joint probability of the observed allelic read counts across replicates
Beta+Rep (mainly used in allele specific expression [134, 170])	Beta-binomial	Calculate the joint probability of the observed allelic read counts across replicates
edgeR [133]	Negative binomial	Normalize according to library size and then sum across replicates

Table 3.1 Five investigated methods for calling ASB events

The abbreviation of each approach is listed in the first column, followed by the statistical distribution (second column) and replicate processing method (third column). See Materials and methods in Chapter3 for more details on each method.

The numbers of called ASB events varied widely across the five methods (Figure 3.1B). Beta+Pool was the most conservative method, only calling 3,566 ASB events; while Binom+Rep was the most permissive method, calling 8,940 ASB events. Methods based on the beta-binomial

distribution were more conservative than those based on the binomial distribution. In general, ASB events called by the conservative methods were called by the permissive methods. For instance, 81.9% of ASB events called by the most conservative method (Beta+Pool) were called by all of the other four methods. The most permissive method (Binom+Rep) called 99.3% of ASB events from the union of the five methods. For the methods with the same statistical distribution, pooling replicates resulted in more conservative calling compared to those that considered replicates independently. The resulting p-values from ASB calling were also highly correlated between the five methods (Figure 3.1B and Figure 3.2). Overall, approaches based on the beta-binomial or binomial distributions correlated more with each other (average spearman correlation coefficient of 0.95) than with edgeR (average spearman correlation coefficient of 0.91). In summary, five evaluated methods mainly differed in their statistical stringency and the resulting p-values from ASB calling were highly correlated.

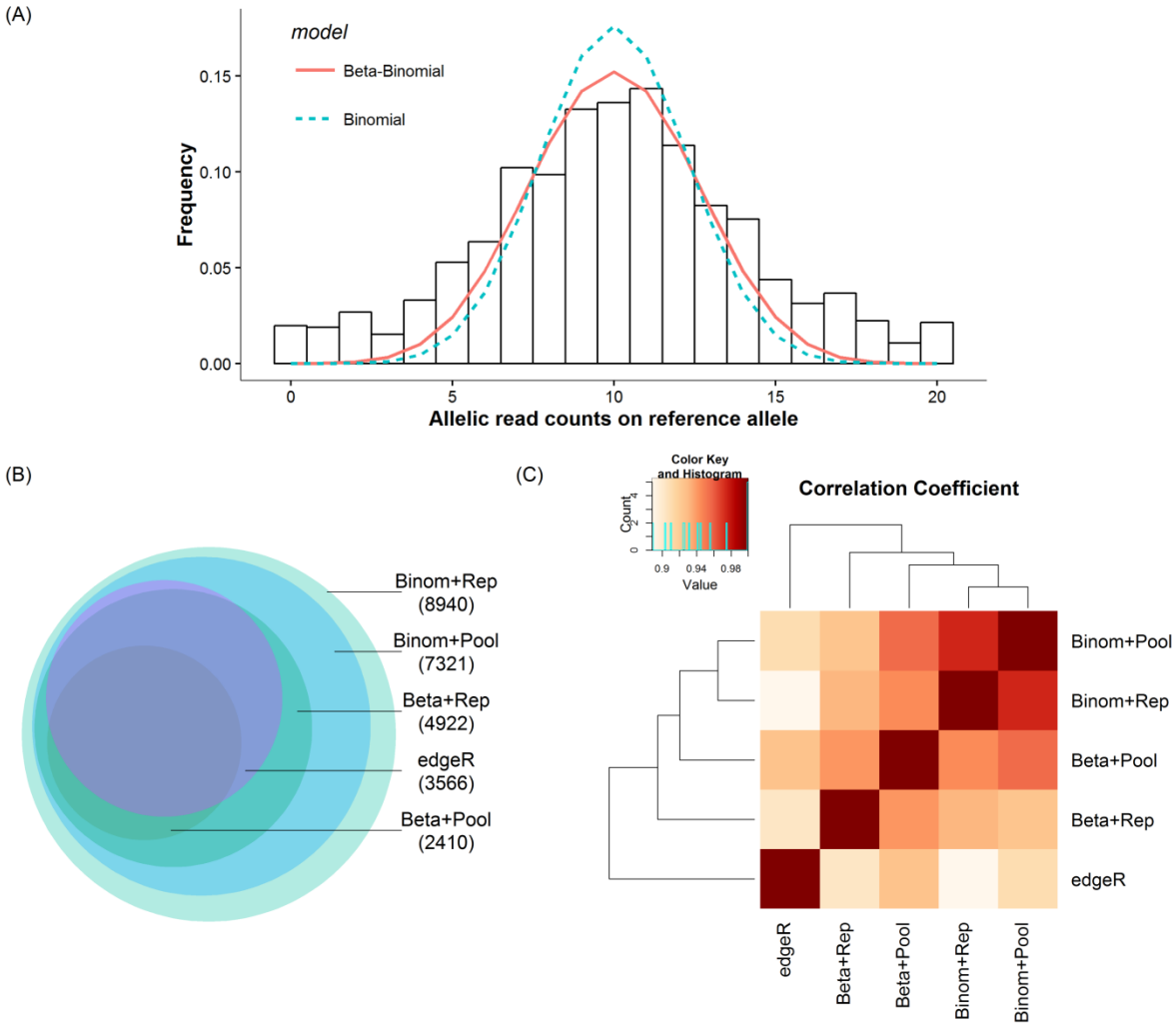


Figure 3.1 Evaluation of five ASB calling methods

(A) The over-dispersion problem of allelic read counts. The histogram indicates the empirical distribution of allelic read counts for the heterozygous sites with read coverage of 20. The fitted curves represent the estimated binomial distribution (blue dash line) and beta-binomial distribution (red line). (B) Venn diagram of the ASB events called by five methods. The number in the parenthesis indicates the number of ASB events called by each method. (C) Averaged correlation of p-values between five methods across 39 TF datasets.

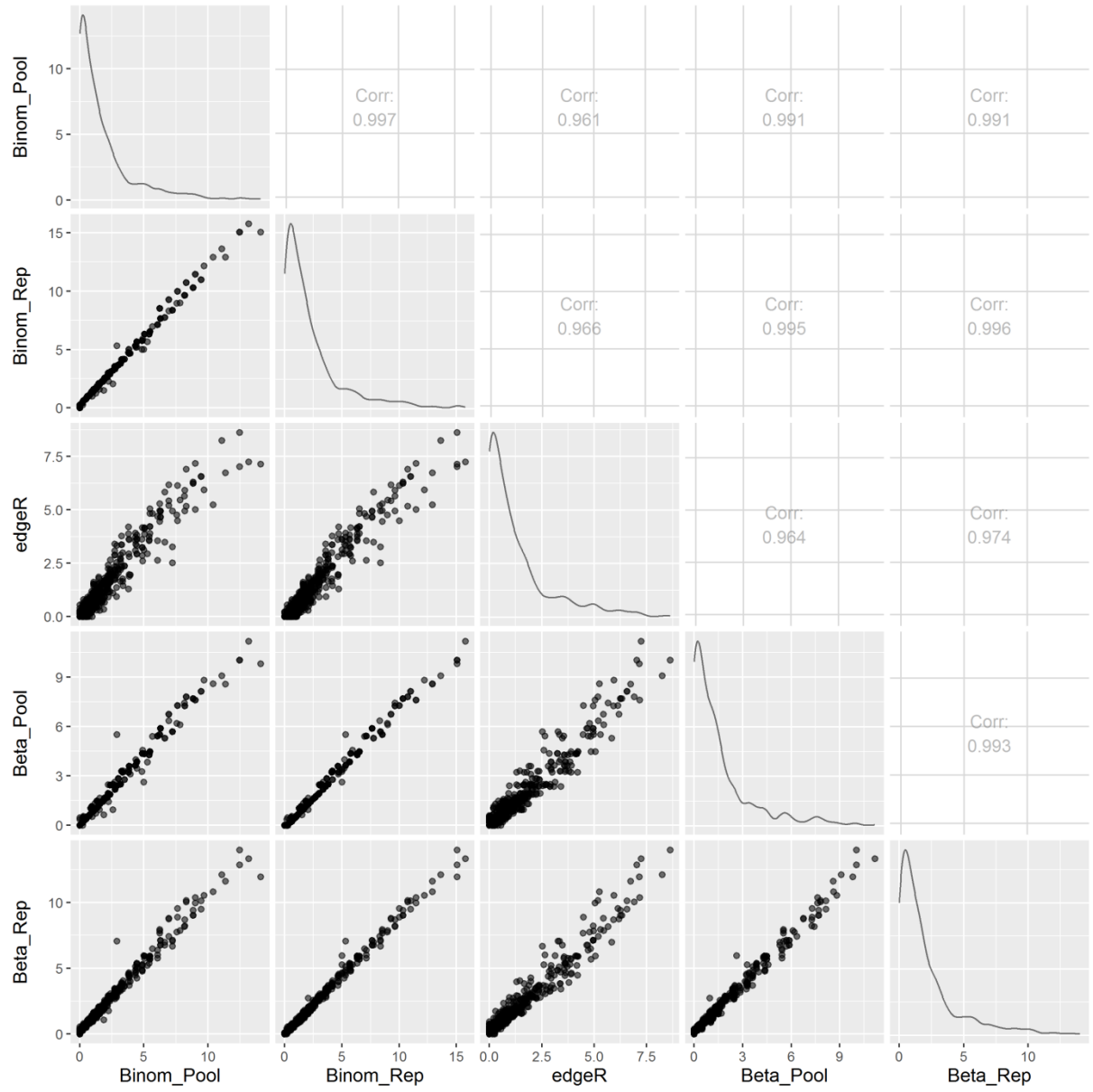


Figure 3.2 Compare the p-values of five ASB calling methods on one dataset

In the scatterplot of the lower panels, each dot represents two p-values in negative log scale for two ASB calling methods at one heterozygous site binding event from EBF1 dataset in GM12878. The diagonal plots show the p-value density of each ASB calling method. "Corr" in the upper panels indicates the correlation coefficient for the symmetric plot in the lower panel.

3.3.2 Traditional binomial-based models show higher allelic DHS correlations than other models

To date, there is no gold standard dataset of ASB events to evaluate ASB calling methods. In this work, we introduce the metric "allelic DHS correlation", which correlates the p-values of ASB calling with DHS allelic imbalances to evaluate ASB calling methods (Figure 3.3A; see Materials and methods in Chapter3). The rationale behind this metric is that the favoured allele in more significant ASB events will show higher allelic imbalance in DHS than in less significant ASB events of similar read coverage.

We evaluated the five ASB calling methods based on allelic DHS correlation. Only 32 out of 39 TF datasets showed significant allelic DHS correlations across the five methods, indicating a limitation of the allelic DHS correlation measure. Across 31 significantly correlated datasets, edgeR showed lower allelic DHS correlation compared to the other four methods (p-value = 5.9×10^{-4} and estimated median difference = -0.010 for Binom+Rep; p-value = 3.0×10^{-2} and estimated median difference = -0.008 for Beta+Pool; p-value = 1.9×10^{-2} and estimated median difference = -0.006 for Beta+Rep; p-value = 4.5×10^{-5} and estimated median difference = -0.013 for Binom+Pool; Wilcoxon signed-rank test). Each of the two binomial-based methods showed higher DHS allelic correlations than the two beta-binomials based methods (p-value = 2.3×10^{-4} and estimated median difference = 0.007 for Binom+Pool and Beta+Pool; p-value = 9.3×10^{-3} and estimated median difference = 0.004 for Binom+Rep and Beta+Rep, one-sided Wilcoxon signed-rank test). Moreover, within the binomial-based methods, pooling replicates together (Binom+Pool), even though it was less permissive, resulted in a higher DHS allelic correlation

(p-value = 2.1×10^{-2}, estimated difference=0.07, Wilcoxon signed-rank test). Overall, the Binom+Pool method delivered the highest performance in terms of allelic DHS correlation (Figure 3.3B), suggesting that the null distribution may be closer to the binomial distribution than a distribution estimated by the beta-binomial distribution.

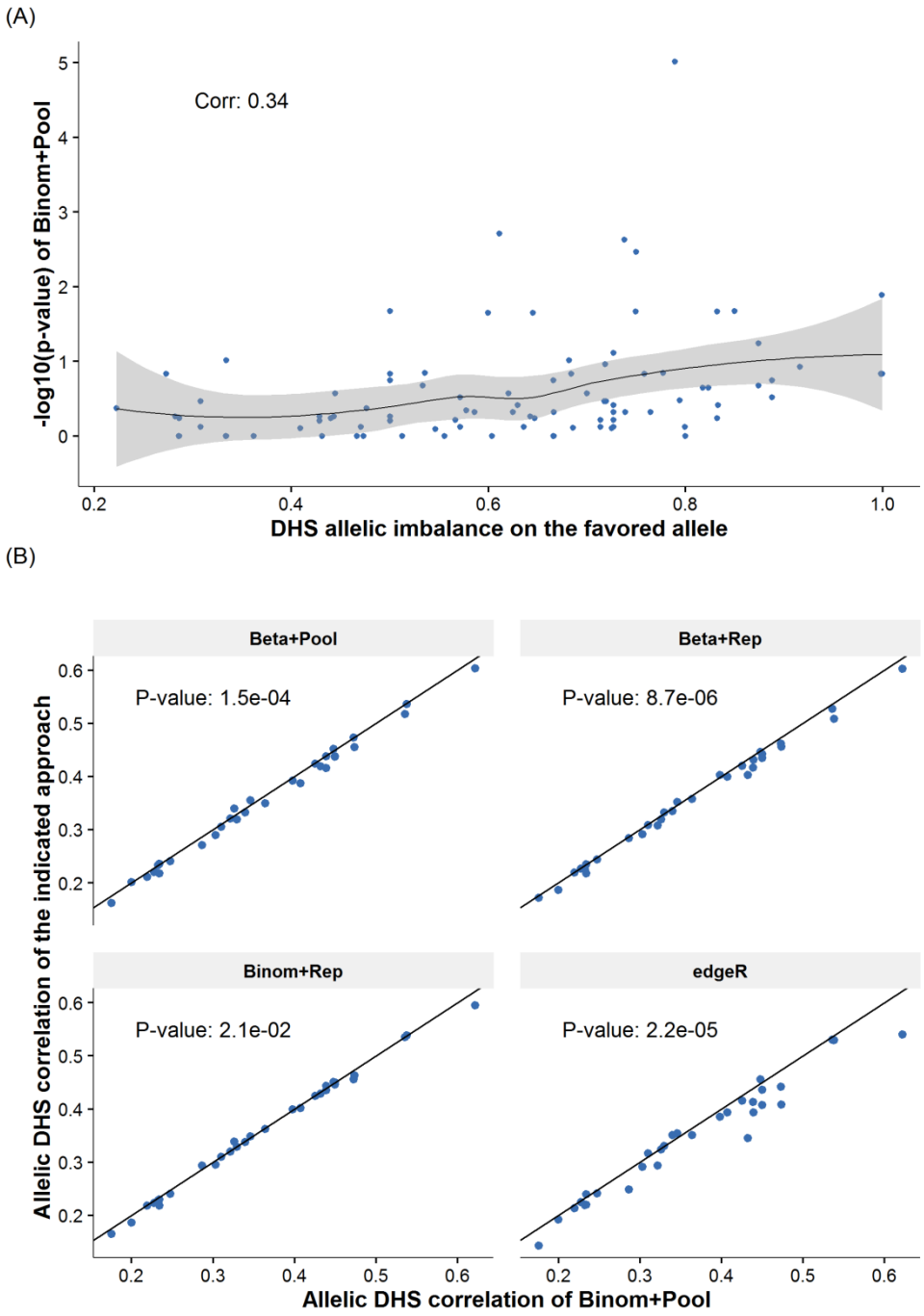


Figure 3.3 Evaluate ASB calling methods based on allelic DHS correlation

(A) An example of allelic DHS correlation. Each point represents the DHS imbalance on the favoured allele (x axis) and the resulting p-value from ASB calling (y-axis) for a heterozygous site in the EBF1 dataset from GM12878

cells. “Corr” indicates the coefficient of spearman correlation. (B) Comparing of Binom+Pool with the other four ASB-calling methods based on allelic DHS imbalance. Each dot represents the allelic DHS correlation coefficient for one TF dataset between the Binom+Pool (x-axis) and the indicated method (y-axis). The provided p-values indicate whether the two methods significantly differ in terms of allelic DHS correlation across the 32 tested TF datasets.

3.3.3 The degree of over-dispersion is overestimated due to mild TFBS alterations

As the binomial based methods perform better without considering over-dispersion, we sought to understand what contributes to the over-dispersion. Previous literature [132, 173, 174] has suggested that the over-dispersion could reflect technical variance or artifacts. We sought to assess if the observed over-dispersion may be more directly related to mild TF binding alterations. We divided non-ASB events (FDR > 0.05 in binomial test) into two groups: 1) TFBS alteration group, in which variants are situated within the best predicted TFBS of the peak and changed the PWM score at least by 0.02; and 2) the remaining non-ASB events. For each TF dataset, the dispersion parameters of the two classes were estimated according to the Beta+Pool method. We found that dispersion parameters were significantly higher in the TFBS alteration group (Figure 3.4; p-value = 1.3×10^{-4}, estimated media difference = 0.07, Wilcoxon signed-rank test), suggesting that the observed over-dispersion in the data is in part due to mild TFBS alterations. Thus, the real null distribution is closer to the binomial distribution than estimated, supporting the use of binomial-based method and potentially explaining the superior performances.

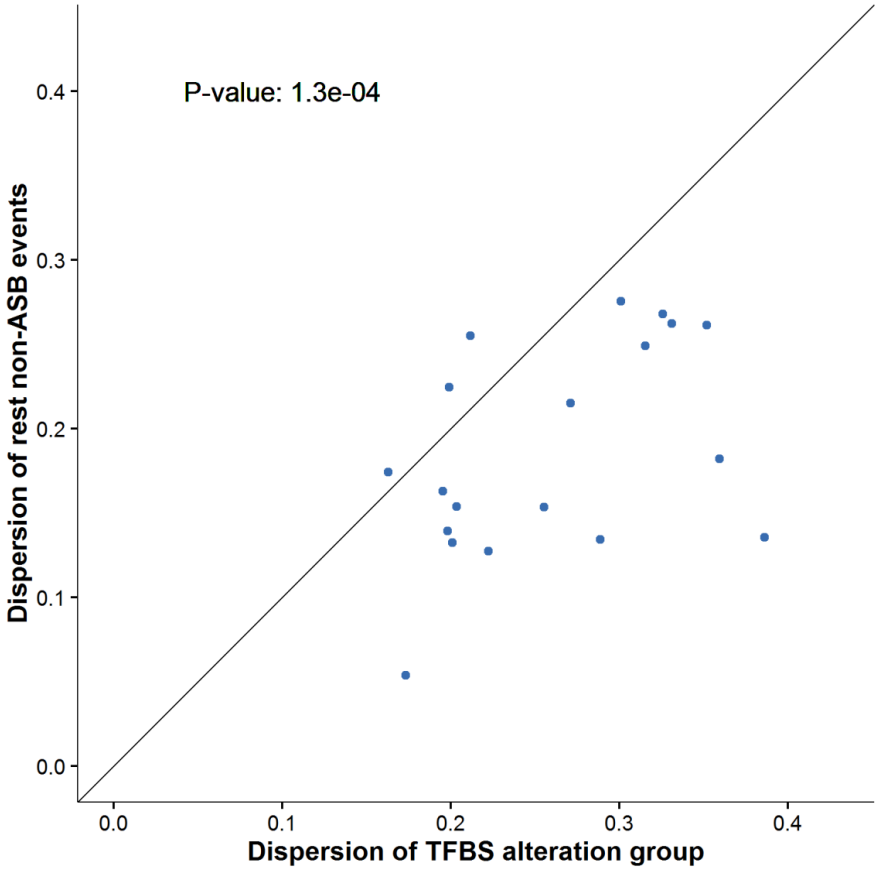


Figure 3.4 TFBS alterations lead to higher degree of over-dispersion in non-ASB events

Each dot represents a TF dataset, and the coordinates indicate the estimated over-dispersion parameters of non-ASB events with TFBS alterations (x-axis) and without TFBS alterations (y-axis). The provided p-value indicates the significance of the test whether the over-dispersion parameters are same for both axes.

3.4 Discussion

In this chapter, we evaluated five ASB calling methods derived from different statistical distributions and replicate processing approaches. Each method called a different number of ASB events based on a statistical threshold, but the ranks of the reported p-values from the ASB calling procedures were highly correlated across five methods. As there is no gold standard ASB

data for benchmarking, we introduced allelic DHS correlation as a measure to evaluate the methods. The Binom+Pool method produced the highest allelic DHS correlation, thus for this performance measure the binomial distribution is the most appropriate for ASB calling. While they have been useful for the study of RNA-seq, beta-binomial based methods appear to overestimate the degree of dispersion for the ASB data used here.

The identification of characteristics of the null distribution is a key challenge for ASB calling as there is no gold standard dataset for non-ASB events. Skelly *et al.* proposed the use of read counts from DNA sequencing data to estimate the degree of over dispersion at heterozygous sites for allele specific expression [173]. This approach requires that the DNA and RNA-seq samples are subjected to the same sequencing and downstream processing procedures. However, this requirement is difficult to meet for most of the ChIP-seq datasets in our project. Alternatively, we might use the control of the ChIP-seq experiment to estimate the degree of over dispersion at heterozygous sites. However, we find that read coverage at heterozygous sites in control experiments is low within TF binding regions (e.g. mean read depth = 0.25 for CTCF in GM12878), such low coverage data might be inappropriate for the estimation of the degree of over dispersion. Future work is needed to create appropriate datasets to estimate the degree of over dispersion in the null distribution.

As the Binom+Pool approach does not address over-dispersion, we explored the potential sources for over-disperson and found that the data suggested a contribution of mild TFBS alteration. The causes of observed over-dispersion has not been deeply explored in ASB events, with one ASB previous study postulating contributions from technical issues, such as sparse read

coverage [132]. As ASB events are associated with genetic or epigenetic alterations, we anticipate that epigenetic differences could also contribute to over-dispersion.

In conclusion, the five ASB calling methods mainly differ in their statistical stringency, but the relative rankings remain highly similar. Until a more reliable gold standard is available to further assess the performance of the methods, we perceive the Binom+Pool method remains the preferred option for ASB calling.

Chapter 4: Predicting the impact of altered TF binding on gene expression based on sequence variants

In Chapter 2 and 3, we compiled ASB events and trained classification models to predict the impact of sequence variants on TF binding. Next we asked whether the altered TF binding events would alter gene expression. In this chapter, we developed regression models to address this question and the developed models further provided mechanistic insights on how the sequence variants alter TF binding and influence gene expression.

4.1 Introduction

Understanding the role of genetic variants in human disease is a fundamental question in medical genetics. Whole genome sequencing has enabled genetics researchers to systematically seek variations that contribute to disease phenotype. Current clinical approaches using DNA sequencing focus primarily on the ~2% of the human genome containing protein-coding exons, as predicting the functional impact of non-coding variants remains a challenge. However, up to 88% of the disease-related variants in genome-wide association studies (GWAS) located in non-coding regions [175]. Non-coding regions are involved in multiple steps of gene regulation, indicating the regulatory roles of non-coding variants on gene expression.

With high-throughput sequencing technology, expression quantitative trait loci (eQTL) studies conduct thousands of single-variant tests to identify the variant associated with gene expression. By overlapping with the genomic annotations, the identified eQTLs were found to be enriched in regulatory regions, e.g. TF-bound regions, suggesting their potential roles in gene regulation [30,

121]. Recently, multiple proximal SNPs have been integrated to predict gene expression levels in regression models [124, 125]. However, causal variants and their functional roles are still challenging to infer due to multiple reasons. First, causal variants are hard to infer in association studies due to the linkage disequilibrium between SNPs [176]. Second, rare variants are not considered in association studies but can lead to various diseases [114, 123]. Third, current approaches separate the SNP identification and functional interpretation into two steps, potentially missing the real functional variants in high linkage disequilibrium.

Both GWAS and eQTL studies highlighted the importance of TF-bound regions, supporting their key roles in gene regulation [30, 121]. TF can bind in a sequence-specific manner to short DNA segments, named TF binding sites (TFBSs) [17]. Outside the core TFBS, flanking regions and binding of partner TFs also contribute to TF binding [52, 177]. The impact of variants on TF binding can be better predicted due to recent progress in machine learning and the availability of ChIP-seq binding data sets. For instance, deep learning models can detect various sequence patterns within broader ChIP-seq peak regions (e.g. 1,000bp), showing superior performance compared with traditional approaches [44, 45]. Alternatively, allele specific binding events, in which a TF significantly prefers one allele over another at heterozygous sites, providing high quality data to investigate the impact of single nucleotide variant in the same cell context [52, 132]. However, the relationship between altered TF binding and gene expression levels is complex. TF binding events can be non-functional towards the expression of nearby genes [178], or the altered TF binding can be buffered by other TF binding events [63]. To further understand the function roles of regulatory variants, a key question is that which altered TF binding events would alter the expression output of the target gene.

Here we present TF2Exp to infer the potential impact of altered TF binding and suggest mechanisms by which regulatory variants act. We trained regression models to predict gene expression by considering altered TF binding events in associated regulatory regions across 358 lymphoblastoid cell lines (LCLs). For 3,060 genes, TF2Exp has suitable predictive performance ($R^2 > 0.05$), revealing 3.7 key altered TF binding events on average for each gene. We found that the selected TF-binding events in promoters showed higher effect sizes than events situated in distal regulatory regions. The TF2Exp models showed comparable performance to SNP-based models in cross validation. Taken together, for a subset of modeled genes, TF2Exp advanced our understanding on the functional roles of non-coding variants on gene expression in terms of altered TF binding.

4.2 Materials and methods

4.2.1 Quantification of gene expression levels from RNA-seq data

RNA-seq and variant calling data for 358 LCLs (individuals) were downloaded from the GEUVADIS project [121] and the 1000 Genomes Project [1] (Appendix C.1). Individuals covered 4 populations, including 89 North-Europeans from Utah (CEU), 92 Finns (FIN), 86 British (GBR) and 91 Toscani (TSI). For each population, we built sex-specific transcriptomes, in which SNP positions with a minor allele frequency (MAF) ≥ 0.05 were replaced by N (representing any of the four nucleotides A, C, G, T) using scripts from [179]. RNA-seq data were processed using Sailfish (version 0.6.3) [180], and the expression level of each gene was quantified as transcripts per million reads. The resulting expression data were normalized via multiple steps, including standardization, variation stabilization, quantile normalization and

batch effects removing (*i.e.* population and gender, and 22 hidden covariates) by PEER [181] (Appendix Figure C1). Genes on sex chromosomes or with near zero variance expression levels were removed, leaving 16,354 genes for model training.

4.2.2 Associating regulatory regions and TF-binding events to genes

For GM12878, we downloaded Hi-C data [179], which measure putative physical interactions between DNA fragments by means of proximity scores. The average resolution of Hi-C data was 3.7Kb [179]. Gene promoter was defined as the ±2Kb region centered at the gene start position (outermost transcript start position in ENSEMBL GRCh37) and flanking regions extended by overlapped Hi-C fragments (interacting DNA fragments in Hi-C data, Figure 4.1). Any other Hi-C fragments within 1Mb of the gene body (region between the outermost transcript start and end) were defined as "distal regulatory regions" of the gene if they contacted the promoter region (proximity score > 0.4) [179]. We downloaded uniformly processed DNase I hypersensitivity (DHS) and ChIP-seq peaks for 78 distinct TFs in GM12878 from the ENCODE project [30]. Because DHS is a general indicator of TF binding [182], we would refer DHS peaks as TF ChIP-seq peaks for convenience. We assumed that each ChIP-seq peak represent one TF binding event. TF binding events were associated to a gene if they overlapped either the promoter or distal regulatory region of a gene.

4.2.3 Predicting sequence variation impact on TF binding events

Variant calling data of each individual was downloaded from the 1000 Genomes project (release 20130502) [1], and we only considered single nucleotide variants and small indels less than 100bp. TF ChIP-seq peaks derived from GM12878 were used as the reference for all studied

individuals. The impact of a variant within a TF binding event was calculated as binding score difference between the altered and reference allele given by the corresponding DeepSEA (v0.93) TF binding model trained from GM12878 data [44]. In order to accommodate the analysis of multiple variants within a TF binding event, we modified DeepSEA code to calculate the binding score of each allele using the 1,100bp region centred at the ChIP-seq peak max position (the original code calculated the scores for the 1,100bp region centered at each variant). TF ChIP-seq peaks with multiple peak max, and overlapped peaks from the same experiment (*e.g.* from the same TF), were split at the center of each two neighbour peak max positions. At heterozygous positions, the binding score difference was multiplied by 0.5. Score differences of multiple variants within the same TF-binding event were summed as the overall alteration of the indicated binding event.

4.2.4 Quantitative models of gene expression

LASSO regression on gene expression: We developed regression models to predict the expression level of a gene using altered TF binding events associated with that gene based on the following equation:

$$Y_i \sim \sum_{k=1}^{n} \beta_k \Delta TF_{i,k} + \epsilon \quad (1)$$

where Y_i is the expression levels of gene *i* across studied individuals, *n* is the number of TF binding events of gene *i* , $\Delta TF_{i,k}$is the alteration of TF binding event *k* across studied individuals and β_k is the effect size of TF binding event *k*. In equation (1), Y_i is the response and $\Delta TF_{i,k}$ is the input feature for the LASSO regression model, which is trained by the glmnet package [183] in R [172]. Model performance was evaluated by 10-fold nested cross-validation, in which the internal folds identified the optimal hyper-parameter lambda, and outer layers tested model

performance. Model performance was measured by R^2 as the square of the correlation between predicted and observed expression levels. The trained models would select a subset of TF binding events as key features of which effect sizes were not zero. When Hi-C proximity scores were used as the prior to select features, the prior (penalty.factor in the glmnet function) was set to "1 – proximity score".

Defining interactions between two TF binding events: For TFs known to interact in the BioGrid database [184], we created interaction terms between pairs of TF binding events (one from each TF) if they satisfied one of the following conditions: 1) two binding events overlapped by at least 200bp; or 2) their regulatory regions were reported to interact in the Hi-C data.

SNP based models/Models with different input feature sets: Following the same procedures as described in the work of Gamazon *et al*. [124], for each gene, we trained regression models based on multiple SNPs to predict the expression level of each gene. The variant calling data of each individual were converted to allelic dosage by plink2 [185]. We only considered SNPs with MAF > 0.05 and within 1Mb of the gene body regions. The regression formula of SNP-based models was as follows:

$$Y_i \sim \sum_{k=1}^{n} \beta_k X_{i,k} + \epsilon$$

Where Y_i is the expression levels of gene *i*, *n* is the number of SNPs, and X_k was the number of alternative alleles of $SNP_{i,k}$.

4.2.5 Gene ontology enrichment analysis for the top performance genes

We used the GREAT web tool (http://bejerano.stanford.edu/great/public/html/) [186] to conduct gene ontology enrichment analysis for the top 400 performance genes with default settings. The gene ontology terms of each gene is obtained from UniProt database (http://www.uniprot.org/uniprot/P78410) [8].

4.2.6 Analyzing selected features using FANTO M5 data

FANTOM5 project associated enhancers to genes based on the activity correlation between the candidate enhancer and the gene promoter across multiple tissues or cell types [94]. For each predictable gene, we counted the overlap between the associated enhancers specified by FAMTOM5 and the selected (or unselected) TF binding events given by TF2Exp models. The overlap statistics were aggregated across all the predictable genes. The overall enrichment of selected features in enhancer regions was given by Fisher's exact test.

4.2.7 External validation

We obtained microarray expression data in LCL from individuals of 3 populations, including CEU, Chinese (CHB), and Japanese (JPT) [187], for external validation of our models. The microarrays included expression data of 47,294 probes, which mapped to 15,997 unique Ensembl genes.

4.2.8 Code and data availability

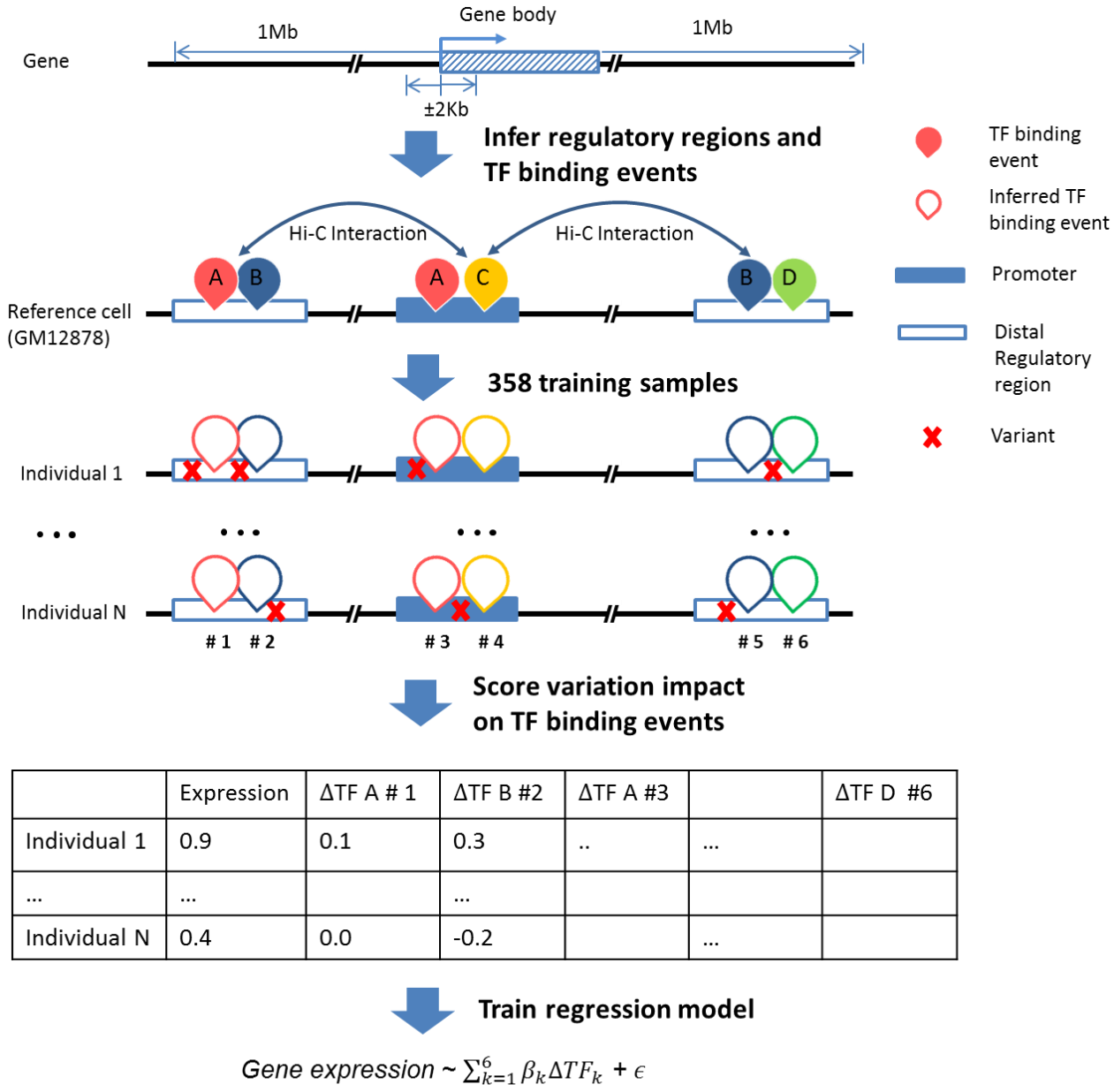
The code and data links used in this work can be found at www.github.com/wqshi/TF2Exp. Multiple packages have been used for data processing and model training, including BEDTools [188], vcftools [189], caret [153] and ggplot2 [190].

4.3 Results

4.3.1 TF2Exp: regression models to predict the impact of altered TF binding on gene expression

We developed TF2Exp, a gene-based computational framework to assess the impact of altered TF binding events on gene expression (Figure 4.1). Variant calling data (single nucleotide variants and small indels) and gene expression data of 358 LCLs were obtained from the 1000 Genomes [191] and GEUVADIS projects [121]. TF binding events of the studied LCLs were inferred based on the available ChIP-seq data of one LCL, GM12878, for which the called ChIP-seq peaks of 78 distinct TFs and DHS datasets were obtained from the ENCODE project [30]. Because DHS is a general indicator of TF binding [182], we referred DHS sites as TF binding events. Gene promoter region was defined as $\pm$2Kb of gene start site and flanking regions extended by overlapped Hi-C fragments based on the Hi-C data from GM12878 [179] (Material and Methods, Figure 4.1). Distal regulatory regions of a gene are the Hi-C fragments interacted with its promoter regions suggested by Hi-C data. TF binding events were associated to a gene if they overlapped either the promoter or distal regulatory region of the gene. The impact of each single variant within a TF binding event was scored using DeepSEA [44], which provided precomputed model for the corresponding TF. The impact of multiple variants within the same TF binding event were summed to generate an overall alteration score of that TF binding event in

each individual. On average, each gene had 420.0 altered TF binding events within 36.6 regulatory regions (promoter and distal regulatory region) across the collected samples. Based on alteration scores of TF binding events in each individual, regression models were trained by LASSO [183] to predict gene expression per individual and select key TF binding events.



	Expression	ΔTF A # 1	ΔTF B #2	ΔTF A #3		ΔTF D #6
Individual 1	0.9	0.1	0.3	..	...	
...	...		...			
Individual N	0.4	0.0	-0.2		...	

Train regression model

$$Gene\ expression \sim \sum_{k=1}^{6} \beta_k \Delta TF_k + \epsilon$$

Figure 4.1 The overview of the TF2Exp framework

(A) Infer regulatory regions and TF binding events of each gene based on the reference cell line (GM12878). Distal regulatory regions were associated to the target gene according to the Hi-C data. All the TF binding events on the promoter or distal regulatory regions of a gene were associated to that gene. (B) Score the alteration of TF binding

events based on the overlapped variants in each individual. (C) Train regression models for each gene across the collected individuals.

4.3.2 The expression of a subset of genes are predictable by TF2Exp

TF2Exp models showed an average performance (R^2) of 0.048 across 15,914 successfully trained models in 10-fold cross validation, with most models having low predictive power (Figure 4.2). To focus on predictable models and genes, we set an R^2 threshold of 0.05 as in [125]. Above that threshold, predictable genes accounted for 19.2% of the investigated genes, a subset upon which we focused in later feature analysis. As in the work of Manor *et al.*[125], we observed a significant correlation between the variance of the gene expression and the TF2Exp performance for the predictable genes (Spearman correlation 0.25, p-value = 4.0×10^{-43}, Appendix Figure C2). The top 400 genes are enriched in the genes related to diseases in immune system, such as graft-versus-host disease, allograft rejection and autoimmune thyroid disease. For example, one of the highest performance gene is BT3A2 (R^2=0.83, Appendix Figure C3), which is associated with T cell mediated immunity and interferon-gamma secretion. To assess the randomness in the model training process, we set up control models in which gene expression was shuffled across individuals but preserving TF binding features. Control models showed an average R^2 of only 1.9e-4 (Figure 4.2), supporting the non-random signal captured by TF2Exp models.

We next sought to determine if additional information could substantially improve model performance. We assessed whether prior knowledge, such as the proximity score in Hi-C data and known TF-TF physical interactions, could improve TF2Exp models. We introduced the proximity score of Hi-C interactions to guide model fitting, so that TF binding events on highly-

interacting regions were less regularized by LASSO (Materials and methods in Chapter4). We found that adding Hi-C proximity score generated little improvement of 9.4×10^{-4} for R^2 on average (p-value = 4.1×10^{-6}, Wilcoxon signed-rank test), suggesting that the original TF2Exp models had captured most of the signal from the highly-interacting regions. We also tested models that included interaction terms for known TF-TF physical interactions (see Materials and methods in Chapter4). Adding TF-TF interactions significantly reduced the model performance by 7.7×10^{-6} on average (p-value < 2.2×10^{-16}, Wilcoxon signed-rank test, Figure 4.2), suggesting that TF-TF interaction terms did not add additional information beyond individual TF binding events. We therefore focused on the original (simpler) TF2Exp models in the next stages of our analysis.

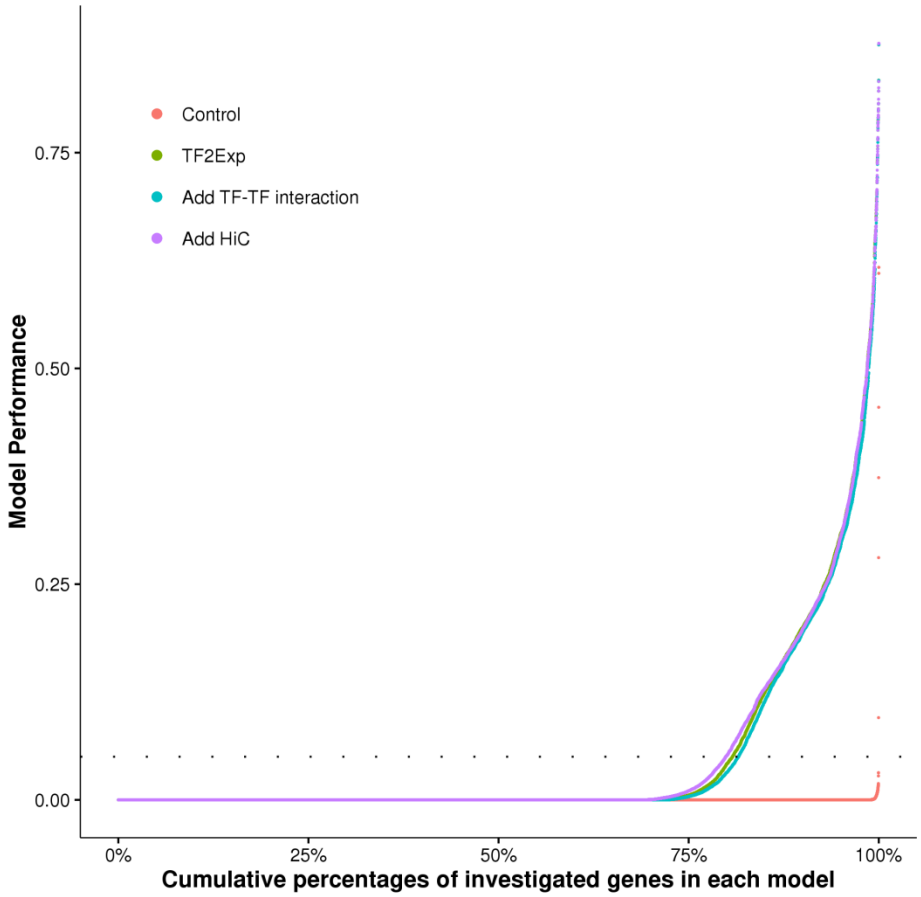


Figure 4.2 Compare the performance of different TF2Exp based models

The performances (R^2) of all the investigated genes (y axis) are plotted in ascending order within each type of TF2Exp models, and x axis represents the cumulative percentage of each gene. The dashed line indicates the defined performance threshold of 0.05 for predictable genes.

4.3.3 Alteration of DHS, RUNX3, and CTCF binding are the most frequently selected features

We sought to identify frequently selected features in TF2Exp models (where a feature was the alteration score of a single TF binding event). On average, TF2Exp models selected 3.7 features (with a minimum of 1 and a maximum of 30) for predictable genes. The top five frequently

selected TFs included DHS, RUNX3, CTCF, EBF1 and PU.1, accounting for 34.2% of the selected features (Appendix Figure C4). DHS indicates accessible chromatin regions [192], and CTCF is a key regulator for chromatin structure [193]. The remaining three TFs perform important roles in tissue-specific regulation in LCL, e.g. RUNX3 for immunity and inflammation [194], EBF1 for B lymphocyte transcriptional network expression [195], and PU.1 for lymphoid development [196]. Consistent with their frequency in selected features, TFs with more genome-wide binding events were selected more often in TF2Exp models (Pearson correlation 0.97, p-value $< 2.2 \times 10^{-16}$).

4.3.4 The contributions of promoter features are greater than distal regulatory regions

We next examined the locations and effect sizes of selected features. We observed significant depletion of selected features in distal regulatory regions compared with promoter regions (p-value $< 2.2 \times 10^{-16}$, odds ratio = 0.32, Fisher's exact test). The selected features in promoters were mostly within 10Kb of gene start positions, while selected features in distal regulatory regions were distributed within ~500Kb. The effect sizes of the selected features decreased rapidly against their distances to gene start sites (Figure 4.3). The selected features in promoter regions exhibited significantly larger absolute effect sizes than in distal regulatory regions (p-value $< 2.2 \times 10^{-16}$, estimated median difference = 0.02, Wilcoxon rank-sum test, Appendix Figure C5). In addition, TF binding events in promoters show more positive effect (59.3%) than in distal regions (54.4%) for gene regulation.

The regulation pairs between selected distal features and genes were supported by other data, as 48.8% of them overlapped with the enhancer-gene pairs specified by the FANTOM5 project [94]

(see Materials and methods in Chapter4). In addition, the selected distal features were significantly enriched in the enhancer regions associated the same gene compared with unselected distal features (p-value = 1.5×10^{-9}, odds ratio = 1.3, Fisher's exact test), supporting the functional role of the selected distal TF binding events.

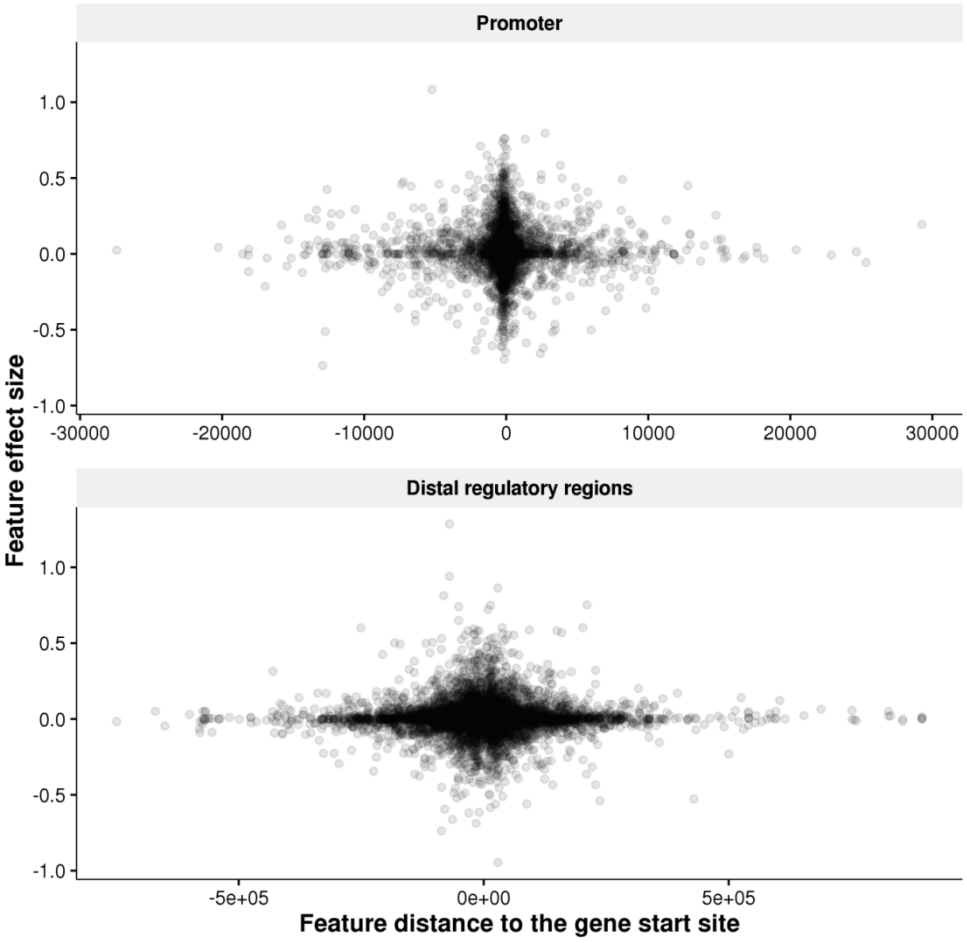


Figure 4.3 The effect sizes of selected features decrease rapidly with their increasing distances to the gene start positions

Each dot represents one selected feature (TF binding event) of predictable genes, and the coordinates indicate the feature distance to gene start site (x axis) and the feature effect size (y axis) obtained in TF2Exp models. The feature effect sizes are plotted separately for promoter regions (top panel) and distal regulatory regions (bottom panel).

4.3.5 TF2Exp models perform comparably to SNP-based expression models

We compared the performance of TF2Exp against existing SNP-based models [124, 125]. TF2Exp models use altered TF binding events as predictors, which combine the impact of multiple variants within a single TF binding event using pre-trained TF binding models, while SNP-based models predict the alteration of gene expression based on the presence of SNPs (e.g. within 1Mb from the gene body) without consideration of potential functional roles (see Materials and methods in Chapter4). We trained both models on the same set of variants (SNPs in all the TF-binding events, SNPinTF), and named two models as TF2Exp-SNPinTF and SNP-SNPinTF. Two models showed comparable performance among the shared predictable genes (p-value = 0.10, Wilcoxon signed-rank test, Figure 4.4). In addition, the default SNP models outperform TF-SNPinTF models with moderate significance (p-value = 0.04, estimated median difference = 8.9×10^{-4}, Wilcoxon signed-rank test), implying that SNPs outside of TF binding events are informative for predicting gene expression.

Next, we compared the selected features in two models. Most of the selected SNPs (66.6%) in the SNP-SNPinTF models overlapped selected TF binding events (74.1%) in TF2Exp-SNPinTF for the same gene. A subset of the selected SNPs (21.1%) overlapped more than one selected TF binding events, revealing multiple roles of SNPs in gene regulation. Only 18.1% of the overlapped SNPs produced the highest impact in the selected TF binding events, highlighting the importance of other SNPs within the same TF binding event. Overall, TF2Exp models simplified the interpretation of SNPs considering their complex roles in multiple TF binding events.

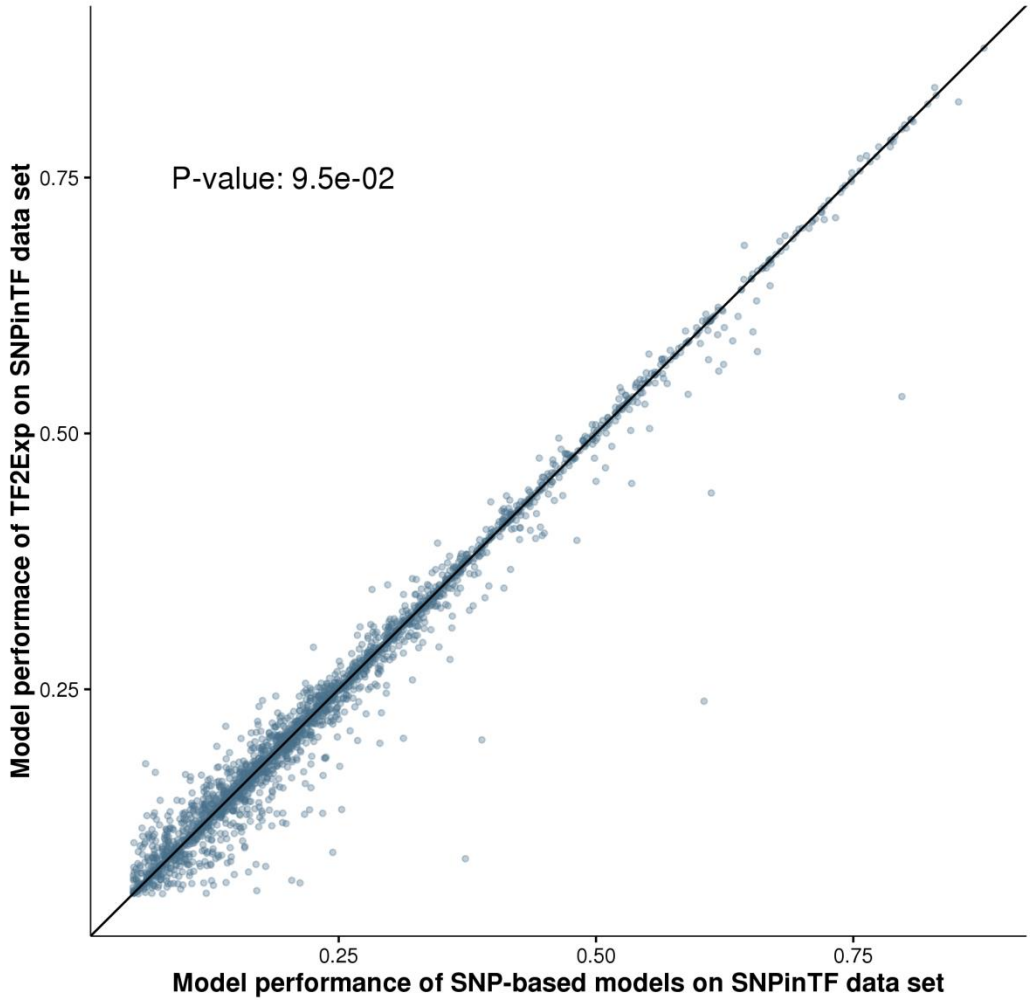


Figure 4.4 TF2Exp models are comparable to SNP-based models

The performance of SNP-based models (x axis) is compared against TF2Exp models (y axis) based on the same SNPinTF datasets (SNPs in the TF binding events). The labeled p-value indicates the significance of different predictive power between two kinds of models given by Wilcoxon signed-rank test.

4.3.6 Uncommon variants improve model performance for a small portion of genes

As TF2Exp models can distinguish the impact of variants in TF-binding events, we investigated the contribution of uncommon (MAF ≤ 0.05) variants to model performance. First, we train

TF2Exp models based only on uncommon variants. The uncommon-variants based models showed much lower average performance (R^2) of 0.011 compared with models based on all the variants (R^2 of 0.048), suggesting the main contribution of common variants in the TF2Exp models. Next we checked the model performance improvement after adding the uncommon variants on top of the common variants. We found that adding uncommon variants only improved a small portion (11.5%) of the models, and the improvements were positively correlated with the performance of uncommon variants models (Pearson correlation coefficient 0.43, p-value $< 2.2 \times 10^{-16}$, Figure 4.5). The improvement can be negative if performances of uncommon variants models were near zero, suggesting the noise caused by the uncommon variants can dilute the information provided by common variants.

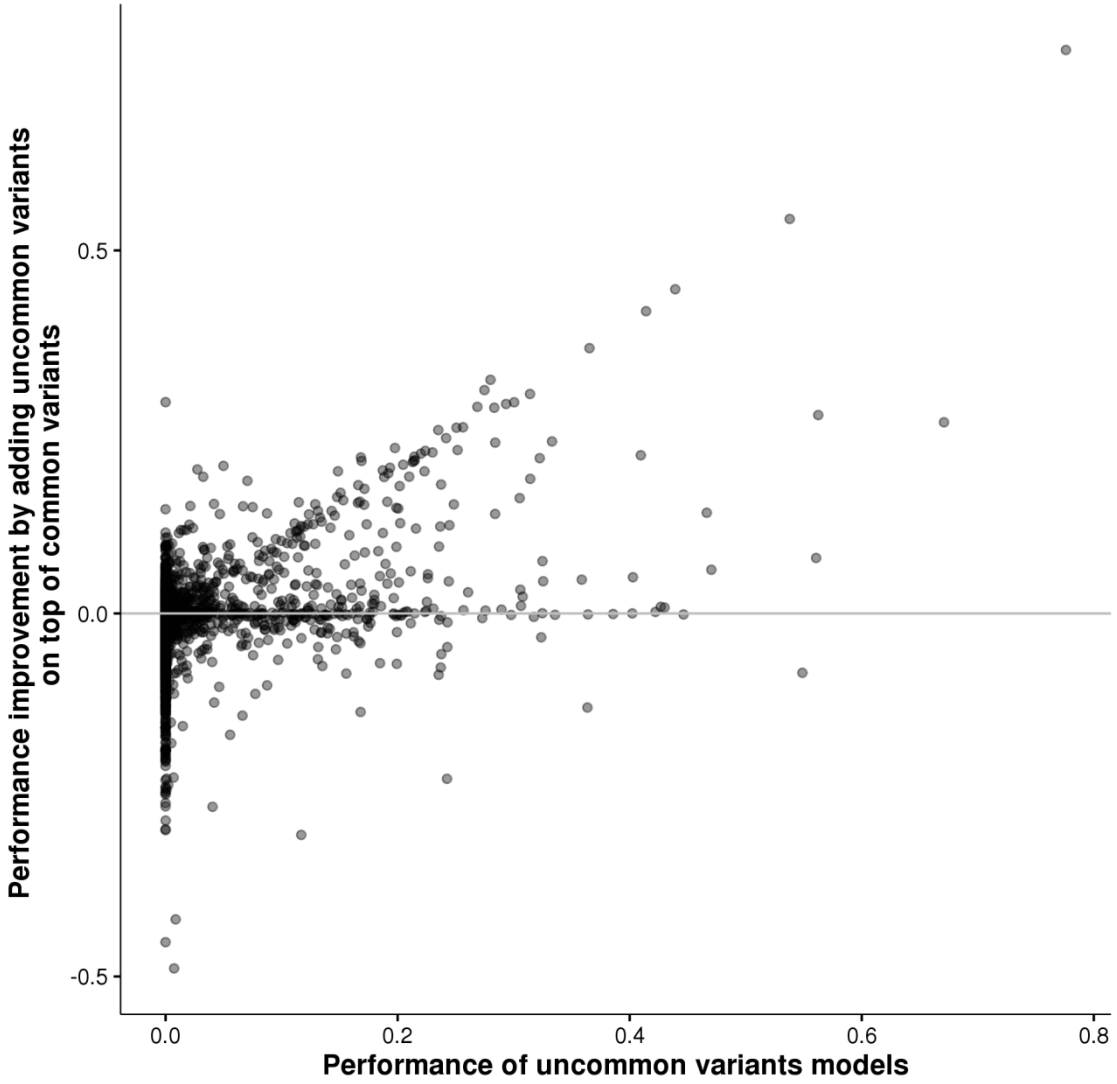


Figure 4.5 Uncommon variants improve the TF2Exp performance for a subset of genes

Each dot represents one predictable gene in the TF2Exp models. The contributions of uncommon variants were measured in two ways: 1) model performance when trained only using uncommon variants (x axis); 2) performance improvement after adding common variants on top of common variants (y axis).

4.3.7 TF2Exp models exhibit robust performance in external validation datasets

We sought to evaluate the models of predictable genes on external datasets. We obtained 256 LCL microarray expression data [187], including 80 CEU, 87 Chinese (CHB) and 89 Japanese (JPT) LCLs (Materials and methods in Chapter4). First, we evaluated the agreement between microarray and RNA-seq data on the 79 shared CEU LCLs in our training data. Relative expression levels across all genes within each individual were concordant between microarray and RNA-seq experiments (average Spearman correlation of 0.76), supporting the overall consistency between the two data sets. However, when we considered a single gene across population, the correlation between the two platforms was low (average Spearman correlation of 0.19). Therefore, we expected models trained on RNA-seq data to have an upper limit performance when applied to microarray data. Next, we used TF2Exp models to predict gene expression levels on the unshared CHB and JPT individuals in the microarray data. TF2Exp models showed an average correlation of 0.17 in CHB and 0.16 in JPT (similar to SNP models, which obtained an average correlation of 0.18 and 0.17 respectively).

An example of a highly-performing gene in the external validation is illustrated in Figure 4.6, comparing the predicted and observed expression. TF2Exp identified 4 contributing TF binding events (Table 4.1), of which two events had greater weights: DHS (chr22:45711760-45711910, effect size: -0.325) and MEF2A (chr22:45771822-45772122, effect size: 0.334). Variant rs104664 in NA12874 increased the score of DHS, predicting lower expression levels of the gene, while rs5765304 in NA11809 increased MEF2A binding scores, resulting in a higher predicted expression.

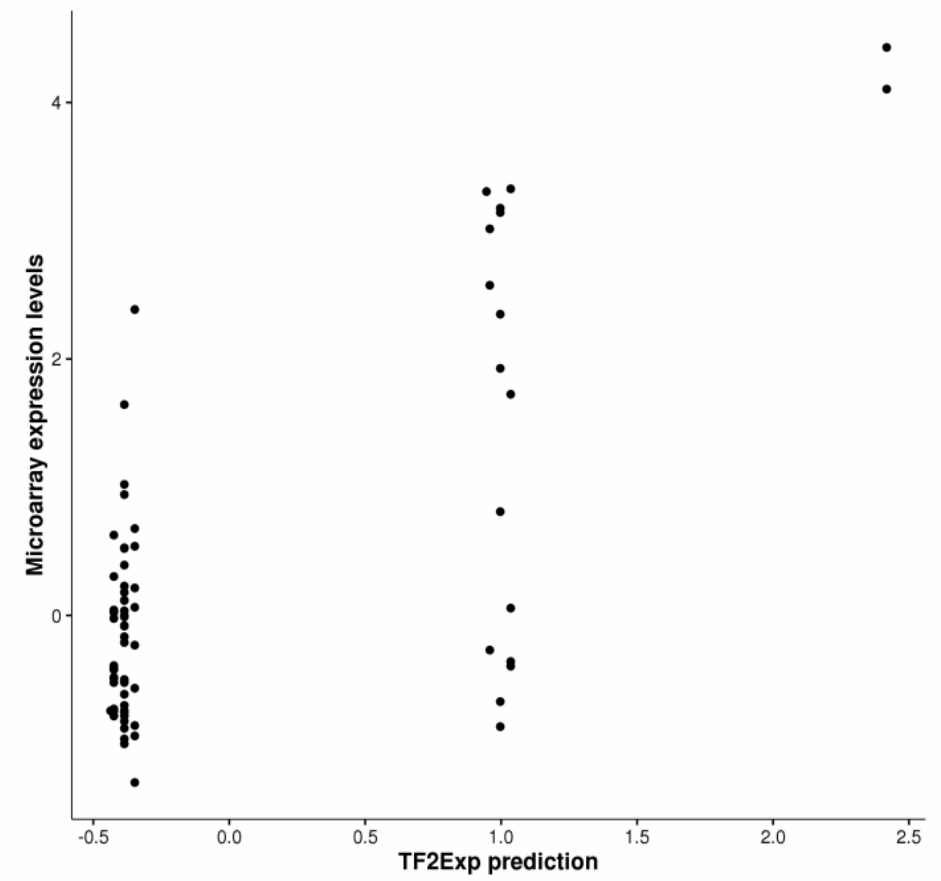


Figure 4.6 Performance of TF2Exp for FAM105 gene in the external validation set

Each point represents one CHB tested individual and its coordinates indicate the predicted expression given by TF2Exp model (x axis) and the observed expression (y axis).

Chr	TF	TF start	TF end	Effect size	Variant	Variant info	MAF	Impact	Max indivi	Min indivi
chr22	PU.1	45697931	45698207	2.31e-02	rs9615099	45698149:T:A	0.344900	-2.62e-3	0|0	0|0
					rs116548001	45698196:G:C	0.015730	-2.10e-2	0|0	0|0
	DHS	45711760	45711910	-3.25e-01	rs104664	45711854:G:A	0.115700	4.13e-2	0|0	1|1
	MEF2A	45771822	45772112	3.34e-01	rs143026443	45771973:C:T	0.002247	-6.19e-3	0|0	0|0
					rs5765304	45771974:G:A	0.113500	9.78e-3	1|1	0|0
	RUNX3	45771818	45772188	1.47e-03	rs143026443	45771973:C:T	0.002247	-1.16e-2	0|0	0|0
					rs5765304	45771974:G:A	0.113500	2.46e-2	1|1	0|0
					rs114074260	45772173:G:T	0.001124	9.59e-3	0|0	0|0

Table 4.1 Selected TF binding events and overlapped variants for FAM118A gene

TF2Exp models can illustrate the contribution of selected features (first four columns) in term of feature effect size (5^{th} column) given by LASSO. The 7^{th} column (variant info) indicates the position, reference allele, and alternative allele of the variants, and 8^{th} column represents the minor allele frequency. The impacts of the variants on the overlapped TF binding events are given in the 9^{th} column. Last two columns indicate the genotypes of the individual with maximum and minimum expression across the population (0 indicates reference allele, and 1 means alternative allele).

4.4 Discussion

Deciphering the functional role of regulatory variants is a critical challenge in the post-sequence era. Here we present TF2Exp to predict the functional impact of sequence variants in transcriptional regulation. Using data from LCL, we developed regression-based models for the expression level of each gene based on sequence alterations within TF binding events in associated regulatory regions. TF2Exp successfully modeled the expression of 3,060 genes, selecting 3.7 altered TF binding events on average. Alterations within DHS and RUNX3 binding events were the most frequently selected features. The known roles of RUNX3 are consistent with the LCL origin of the expression data. The selected TF binding events within promoters obtained greater weights in the models than the events situated in distal regulatory regions. The TF2Exp models showed comparable performance to SNP-based models, and inclusion of uncommon variants can improve TF2Exp performance for small portion of genes (11.5%). Importantly, the TF2Exp models provide mechanistic insights into how non-coding variants influence gene expression.

Illuminating the functional roles of regulatory variants is a critical challenge for the interpretation of whole genome sequence data, which has motivated the creation of predictive models for gene expression based on DNA sequence. Several SNP-based approaches show potential to predict gene expression based on DNA sequence [124, 125]. However, such SNP-based approaches have limited utility for the inference of causal or functional alterations, because the selected SNPs are usually strongly linked to other SNPs. By focusing on TF binding events as the functional unit in our model, we can evaluate all alterations within the TF binding events regardless of the linkage between SNPs. TF2Exp models determine the relative

importance of each TF binding event for the target gene, and further enable us to interpret the mechanistic impact of the variant within the TF binding event.

Rare variants (MAF < 0.01) can cause large effects in human disease, but their impacts are hard to detect in association-based approaches like GWAS and QTL due to their infrequency [123]. Population analysis found that genes with outlier expression showed an enrichment of rare variants [197]. Such rare variants were strongly enriched in conserved promoter regions and weakly enriched in enhancers and TF binding sites [197]. Knowledge on the functional effects of the variants can help, such as whether a variant is within conserved regions, near a splicing site, or disrupt a TF binding site [114]. TF2Exp models can identify the impact of rare variants within key TF binding events. We found that considering rare variants only improved the performance of a small portion of genes. In the future, models that favor uncommon variants with large effect sizes should be explored.

The predictive performance of the TF2Exp regression models is limited, showing utility for a subset (19.2%) of genes. The inadequate performance might be attributable to several reasons. First, the variance of gene expression attributed to common variants is quite low (e.g. 15.3% estimated in [124]), suggesting that models restricted to TF binding events could only account for a portion of variance in the gene expression. Second, TF2Exp models were limited to the available ChIP-seq datasets of 78 TFs in LCL cell lines, while the majority of human TFs (~1,500 TFs) are uncharacterized. The variant impacts on TF binding events were evaluated by the trained DeepSEA models, of which future improvements could benefit the TF2Exp performance. Third, TF2Exp models focused on TF binding events in transcriptional regulation,

and including post-transcriptional regulation components (e.g. splicing or microRNA-mediated regulation) might explain additional portion of variance for gene expression. Fourth, TF2Exp models are likely to be constrained by the small number of available training samples as adding new features (e.g. TF-TF interaction) decreased the model performance. We expect that new large reference transcriptome data will provide more samples for modeling, including family-focused data that will allow greater clarity about the roles of rare variants. Alternatively, inaccurate prior knowledge could decrease performance of TF2Exp models, for example, TF-TF interactions supported by only single publication may result in elevated false positive interactions in BioGrid database [184].

Identifying the impact of *cis*-regulatory variants on gene expression is critical for understanding the genetic mechanisms contributing to diseases. TF2Exp models are able to predict the impact of altered TF binding on gene expression and provide mechanistic roles of selected TF-binding events and *cis*-regulatory variants. Future enlarged omics data in other cell types will greatly expand the application scope of TF2Exp models.

Chapter 5: Conclusion

In the past decade, high throughput sequencing technologies have generated rich datasets allowing annotation of regulatory regions in the human genome, such as TF binding and open chromatin regions. As our understanding of the genome has increased, it has been noted that disease-associated variants arising from GWAS studies are enriched within these regulatory regions [198]. However, our understanding of regulatory variants remains limited, and interpretation of the functional impact of these variants is therefore a major challenge in current genetics research. This thesis focuses on an important subset of regulatory variants that might alter the binding of TFs. Through the thesis, we have

- compiled a high-quality collection of DNA sequence variations associated with disrupted TF binding (ASB events), a collection that can serve as a gold standard for diverse research questions;
- designed and trained models to predict the variant impact on TF binding;
- evaluated five ASB calling methods;
- developed algorithms to predict the impact of altered TF binding on gene expression.

In this section, I will discuss the contributions of this thesis and how the findings and resources can be used in the near-term within the field. Lastly, I will discuss the broader research directions and opportunities that will contribute to future clinical genome analysis and healthcare.

5.1 Predicting variant impact on TF binding

In order to assess the impact of sequence variation on TF binding, it is essential to define a set of reliable cases in which a subtle variation has a quantitative impact on TF binding. We compiled ASB and non-ASB events from 45 ChIP-seq experiments, greatly expanding the set of variants associated with altered TF binding for the community. Unlike other studies focusing on ASB events [132, 160], we also included non-ASB events in analysis as non-ASB events provide insights into the variants with little impact on TF binding. Multiple chromatin properties were associated with ASB events at allelic level, including DHS and several histone modifications, supporting the coordinated effects between TF binding and chromatin properties [48]. We developed a novel framework to predict variant impact on TF binding, allowing classification between ASB and non-ASB events. The trained models based on sequence features achieved comparable performance with state-of-the-art algorithm deltaSVM [40]. Building on the recognition of DHS data as a general indicator for TF binding [57], we incorporated these experimental properties into the model and achieved a significant improvement in ASB classification.

The ASB data can, and hopefully will, be used broadly within the field to better understand TF-DNA interactions. In considering the distributions of ASB variations across and proximal to TFBS, we found that only 28.7% of the variants in ASB events were situated within TFBS motifs (or comotifs for cooperatively acting TFs). This is consistent with the idea that functional alterations outside the core motif can have importance for TF binding [46]. Though the core motif is critical for TF binding, flanking sequences of core motifs are increasingly a focus of research exploring the relationship between protein structure and TF-DNA binding [177]. Some

of the observations about these flanking regions include GC content preferences within the 10 bp on each side of a core motif for some TFs [199]. Alterations of the sequence 1-2 bp adjacent of the Pho4 motif have been demonstrated to alter the transcription rate of the targeted gene [200]. A potential explanation for the contribution of flanking sequences is related to the topology (shape) of DNA. TFs in different families prefer distinct patterns of DNA shape features [199], and the consideration of DNA shape can improve TFBS prediction both *in vitro* [166] and *in vivo* [201]. Future work could use ASB events to assess the contribution of flanking sequences and DNA shape in altered TF binding.

Within a collaborative project, allelic binding data was used to investigate TFs potentially relevant to X-inactivation [202]. Within the nucleus of female cells, one copy of the X chromosome is silenced by X-chromosome inactivation to compensate for gene dosage. However, some genes escape the X-chromosome inactivation and are expressed from both copies. To identify TFs which may be key for escaping X-inactivation, enrichment analysis of TF motifs proximal to escapee transcription start sites was performed. As the YY1 TF emerged as a candidate, we assessed allelic binding data of YY1 across the X chromosome and determined that YY1 showed bi-allelic binding around bi-allelically transcribed genes, supporting the regulatory roles of YY1 in X-inactivation. This study demonstrates how access to high quality ASB data can be used to better understand regulatory mechanisms.

An alternative approach for ASB classification is to model the whole continuum of allelic imbalance using regression models. It is important to recognize that ChIP-seq experiments produce a continuous spectrum of allelic imbalance, ranging from extreme unbalanced ASB

events to balanced non-ASB events. As we were focused on disrupted TF binding in Chapter 2, we discretized the continuous spectrum into ASB and non-ASB events, and then constructed a binary classifier to predict disruptive variants for TF binding. Compared with discretization approach, there are both advantages and challenges for regression approach. It circumvents potential biases arising from read coverage at heterozygous sites in ASB calling. The statistical significance of an ASB event is determined by a combination of the total read coverage at a position, and the imbalance between observations of the two alleles. The dependence on read coverage might cause inconsistent ASB calling, for instance, an ASB event called in a deeply sequenced ChIP-seq experiment might be classified as a non-ASB event with shallow depth. In the continuous perspective, regression models based on allelic imbalance are more independent of sequencing depth than ASB classification models. A challenge in implementing regression models is caused by imbalanced training data, as the majority of heterozygous sites in the DNA showed balanced binding between the two alleles. Such imbalances favor the majority cases (balanced binding), and the predictions thus are not optimal for detecting the minority class (unbalanced binding). Though multiple imbalanced learning techniques have been developed for classification problems [155] (e.g. up and down sampling), imbalanced learning for regression models is less mature in machine learning. Further regression models could incorporate imbalanced learning techniques used in machine learning methods for ASB analysis.

Prioritizing regulatory variants may be improved by consideration of the variant buffer effect. Studies [48, 62, 203] have demonstrated that if one TFBS is disrupted by a variant, an alternative TFBS in proximity can be bound by the same TF to deliver the similar pattern of expression to the target promoter(s). Such buffer effect can involve homotypic clusters of TFBSs (**HCT**),

which are clusters of TFBS expected to be bound by the same TF within same region. HCT is a common feature in the human genome, with more than half of human gene promoters exhibiting at least one HCT [203]. Conceptually HCT provides several mechanism benefits, including buffer effect for variants, high-affinity cooperative binding and lateral diffusion of TF binding along a DNA segment [63, 177, 203]. HCT is particularly enriched in the promoters of transcription factor genes [203], which suggests functional importance of HCT in gene regulation. Our ASB classification models have explicitly incorporated the number of TFBSs within each ChIP-seq peak as an input feature, while other methods indirectly account for the contribution of HCT in TF binding, such as a *k*-mer approach [37] and deep learning approach [44]. To provide greater mechanistic insights, future ASB calling methods can explicitly consider the buffer effect.

5.2 Evaluating five statistical methods to call ASB events

We have assessed five ASB calling methods based on different underlying statistical distributions and replicate processing approaches. Not surprisingly, we found that the choice of ASB calling methods will greatly impact on the number of identified ASB events. Importantly, the methods produce highly consistent rank orders, which mean that the methods will produce similar results when thresholds are adjusted to produce similar numbers of calls. Among five methods, we recommend the most widely used approach, binomial distribution coupled with replicates pooling, for ASB calling based on the metric of allelic DHS correlation. The recommendation of binomial distribution is further supported by the possibility that over-dispersion in allelic read counts could arise from mild TFBS alterations.

Future studies are needed to determine whether this ASB-specific recommendation is applicable to allele specific expression (ASE) studies. The over-dispersion problem is well known in ASE studies, and the beta-binomial distribution has been widely used to correct for this problem [134, 170]. Over-dispersion in ASE is more difficult to attribute to a specific cause, compared to ASB. It is possible that a portion of the over-dispersion of ASE could be caused by ASB events, which could be a focus for future studies.

5.3 Predicting the impact of altered TF binding on gene expression based on *cis*-regulatory variants

In Chapter 4, we explored which altered TF binding event would impact downstream gene expression. We developed TF2Exp, the first framework (to our knowledge) to predict the expression level of each gene by considering altered TF binding events based on sequence variants. Previous approaches could predict gene expression based on nearby SNPs [124, 125], but the functional roles of the incorporated SNPs were undefined. TF2Exp achieved similar prediction accuracy compared with previous approaches, but provided mechanistic insights as to how the non-coding variants altered TF binding and gene expression. SNP-based models generally lack sufficient statistical power to detect the impact of rare variants, which underlie most familial genetic disorders [114]. In contrast, TF2Exp provides predictions about the impact of rare variants, suggesting that TF2Exp or similar approaches will have wider utility for human genetic studies in the future.

TF2Exp offers a new way to identify the targeted genes of TFs. Previously, targeted genes of a TF has been inferred based on the location and strength of TF binding events [204, 205]. However, TF binding proximal to a gene does not ensure a regulation relationship [63]. For instance, across a set of TFs, silencing the expression of a TF will only impact the expression of a small portion of genes bound by that TF within 10kb of the TSS, suggesting that most TF binding events were not essential (or possibly even not functional) [178]. In TF2Exp, we used three criteria to identify the TF binding events of a target gene: 1) TF binding events located within 1MB of TSS for the targeted gene; 2) chromatin interaction data (Hi-C) supporting the interaction between a TF bound region and the promoter region; and 3) alterations of TF binding events were associated with the expression changes of the target gene. By bringing together chromatin interaction, TF binding, expression and genotype data, we bring a focus upon variants that are more interpretable.

The TF2Exp models were trained based on available tissue-specific data in LCLs, including TF binding and gene expression. The generalization of trained models to other tissues remains to be explored. In previous reports, SNP-based models showed decreased predictive performance when validating trained models against other tissues [124]. We anticipate that TF2Exp models will be more likely to be tissue specific than SNP-based models, as the key TF binding events in TF2Exp models are specific to gene regulation in B-cells. As LCL has been considered as surrogate cells in the studies of primary B cells [206] and neurological disorders [207], TF2Exp models can be applied and tested in such tissues, but the utility of the models in more diverse tissues will be limited. The methods demonstrated within this thesis in building the TF2Exp

models, however, are general and should perform equally well when applied to similar data collections for other tissues.

To apply TF2Exp models more broadly, we need multiple training datasets in the targeted tissue, including matched WGS and RNA-seq of multiple individuals, and binding regions of multiple TFs. Matched WGS and RNA-seq data across large number of individuals are accessible in multiple genomic projects, such as the GTEx project [208], the BLUEPRINT project [209] and TOPMed program (https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed). However, at this time, extensive TF binding data are only available in several tissues or cell lines from the ENCODE project [30], which limits the application of TF2Exp models. Though we anticipate an increasing amount of ChIP-seq data in the future, a practical and more immediate solution could be computationally predicting TF-bound regions. To improve the TFBS prediction accuracy, multiple tools are able to incorporate DNA sequence and chromatin property data (e.g. DHS and histone modifications) of the target tissue [57, 210], and chromatin property data are available for various primary tissues from the ROADMAP project [211]. Thus, the search for functional regulatory variants with TF2Exp may be practical in the near future.

Recent single cell RNA-seq data have shown extensive cell-to-cell variation in transcriptome even within genetically homogenous cell population, including expression levels and expression of specific transcript isoforms [212]. Such studies have highlighted that within populations of cells there may be cells in a variety of maturity states [212]. TF2Exp models are based on the traditional RNA-seq which might miss key characteristics of gene expression by averaging

expression levels across thousands of cells. The availability of cell-specific measurements will allow future development of TF2Exp (and related tools) to explore the relationship between genetic variation and the expression levels of individual isoforms (e.g. alternative promoters) and to assess if the variants influence the variance of expression at single cell level.

5.4 Applications in future healthcare

Affordable DNA sequencing technology is starting to transform healthcare by providing precise genetic information for each patient. The work in this thesis will facilitate future healthcare in two ways.

The compiled data and algorithms in this thesis will facilitate future clinical genome analysis of regulatory regions. DNA sequencing has been widely used in clinical research for genetic disorders, and studies achieved modest diagnostic success (27-73%) depending on the investigated diseases and enrolment criteria of patients [213-215]. As current clinical analysis mostly focuses on variants in protein coding sequences, it is reasonable to assume that a subset of undiagnosed patients might have disorders caused by regulatory variants. The compiled ASB datasets and variant interpreting algorithms in this thesis allow for the identification of candidate variants in regulatory regions for the undiagnosed cases (where WGS data is available). In addition, the mechanistic insights revealed by the identified regulatory variants will be informative to understand the mechanism of the disease. As alterations in regulatory sites often suggest roles for upstream pathways, knowledge of regulatory variants may facilitate the design of personalised therapy for patients, with many therapies potentially tied to existing drugs or other treatments.

An interesting application area for the work of this thesis is the prediction of disease risk for healthy individuals based on WGS data. Companies like 23andMe (www.23andme.com) already provide personal genome services to interpret the disease risk of customers based on genomic testing. At present these services are limited to known high disease-risk loci, such as mutations in the BRCA1 gene for hereditary breast cancer [216]. Though the tested loci in the commercial services are far from complete compared with WGS, personal genome services represent the future direction for individuals to learn about and explore their DNA. Based on WGS data and pathway enrichment analysis, a recent study demonstrated that genes associated with disrupted conserved TFBS are predictive for medical history [217]. Similarly, we anticipate that integrating the interrupted genes predicted by TF2Exp might be able to predict future disease risk. While extensive research remains to be performed, which remains constrained by access to broad WGS and patient history data, the analysis of regulatory alterations offers great promise. Future advances may ultimately provide individuals with personalised guidance for diet, lifestyle and other preventative strategies.

The work of this thesis is enabled by recent advances in multiple fields, such as next generation sequencing, chromatin biology, and machine learning. We believe that future union of multiple fields (e.g. genetics, medicine, and informatics) will create unprecedented possibilities to uncover the genetic mechanisms in human diseases and ultimately improve human healthcare.

Bibliography

1. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA *et al*: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
2. Kulzer JR, Stitzel ML, Morken MA, Huyghe JR, Fuchsberger C, Kuusisto J, Laakso M, Boehnke M, Collins FS, Mohlke KL: **A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell**. *American journal of human genetics* 2014, **94**(2):186-197.
3. Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, Yan C, Khalid O, Kantoff P, Oh W *et al*: **Functional enhancers at the gene-poor 8q24 cancer-linked locus**. *PLoS genetics* 2009, **5**(8):e1000597.
4. French JD, Ghoussaini M, Edwards SL, Meyer KB, Michailidou K, Ahmed S, Khan S, Maranian MJ, O'Reilly M, Hillman KM *et al*: **Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers**. *American journal of human genetics* 2013, **92**(4):489-503.
5. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymaki I *et al*: **The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling**. *Nature genetics* 2009, **41**(8):885-890.
6. Wang S, Wu S, Meng Q, Li X, Zhang J, Chen R, Wang M: **FAS rs2234767 and rs1800682 polymorphisms jointly contributed to risk of colorectal cancer by affecting SP1/STAT1 complex recruitment to chromatin**. *Scientific reports* 2016, **6**:19229.
7. Dodd AW, Syddall CM, Loughlin J: **A rare variant in the osteoarthritis-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate GDF5 expression**. *European journal of human genetics : EJHG* 2013, **21**(5):517-521.
8. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J *et al*: **ClinVar: public archive of interpretations of clinically relevant variants**. *Nucleic acids research* 2016, **44**(D1):D862-868.
9. Li MJ, Wang LY, Xia Z, Sham PC, Wang J: **GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications**. *Nucleic acids research* 2013, **41**(Web Server issue):W150-158.
10. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M: **Linking disease associations with regulatory information in the human genome**. *Genome research* 2012, **22**(9):1748-1759.
11. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J *et al*: **Systematic localization of common disease-associated variation in regulatory DNA**. *Science* 2012, **337**(6099):1190-1195.
12. Reijnen MJ, Sladek FM, Bertina RM, Reitsma PH: **Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden**. *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**(14):6300-6303.
13. De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P *et al*: **A regulatory SNP causes a human**

genetic disease by creating a new transcriptional promoter. *Science* 2006, **312**(5777):1215-1217.
14. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C *et al*: **Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip**. *Nature genetics* 2008, **40**(11):1341-1347.
15. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V *et al*: **FTO Obesity Variant Circuitry and Adipocyte Browning in Humans**. *The New England journal of medicine* 2015, **373**(10):895-907.
16. Smallwood A, Ren B: **Genome organization and long-range regulation of gene expression by enhancers**. *Current opinion in cell biology* 2013, **25**(3):387-394.
17. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements**. *Nature reviews Genetics* 2004, **5**(4):276-287.
18. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities**. *Nature biotechnology* 2006, **24**(11):1429-1435.
19. Meng X, Brodsky MH, Wolfe SA: **A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors**. *Nature biotechnology* 2005, **23**(8):988-994.
20. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpaa MJ *et al*: **Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities**. *Genome research* 2010, **20**(6):861-873.
21. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions**. *Science* 2007, **316**(5830):1497-1502.
22. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J: **Practical guidelines for the comprehensive analysis of ChIP-seq data**. *PLoS computational biology* 2013, **9**(11):e1003326.
23. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome biology* 2009, **10**(3):R25.
24. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.
25. Liu T: **Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells**. *Methods in molecular biology* 2014, **1150**:81-95.
26. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities**. *Molecular cell* 2010, **38**(4):576-589.
27. Mathelier A, Shi W, Wasserman WW: **Identification of altered cis-regulatory elements in human disease**. *Trends in genetics : TIG* 2015, **31**(2):67-76.
28. He Q, Johnston J, Zeitlinger J: **ChIP-nexus enables improved detection of in vivo transcription factor binding footprints**. *Nature biotechnology* 2015, **33**(4):395-401.

29. Rhee HS, Pugh BF: **ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy**. *Current protocols in molecular biology* 2012, **Chapter 21**:Unit 21 24.
30. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.
31. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**(1):16-23.
32. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R *et al*: **JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles**. *Nucleic acids research* 2016, **44**(D1):D110-115.
33. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ: **HOCOMOCO: a comprehensive collection of human transcription factor binding sites models**. *Nucleic acids research* 2013, **41**(Database issue):D195-202.
34. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K *et al*: **Determination and inference of eukaryotic transcription factor sequence specificity**. *Cell* 2014, **158**(6):1431-1443.
35. Mathelier A, Wasserman WW: **The next generation of transcription factor binding site prediction**. *PLoS computational biology* 2013, **9**(9):e1003214.
36. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: **Identification of transcription factor binding sites with variable-order Bayesian networks**. *Bioinformatics* 2005, **21**(11):2657-2666.
37. Ghandi M, Lee D, Mohammad-Noori M, Beer MA: **Enhanced regulatory sequence prediction using gapped k-mer features**. *PLoS computational biology* 2014, **10**(7):e1003711.
38. Huang D, Ovcharenko I: **Identifying causal regulatory SNPs in ChIP-seq enhancers**. *Nucleic acids research* 2015, **43**(1):225-236.
39. Deplancke B, Alpern D, Gardeux V: **The Genetics of Transcription Factor DNA Binding Variation**. *Cell* 2016, **166**(3):538-554.
40. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA: **A method to predict the impact of regulatory variants from DNA sequence**. *Nature genetics* 2015, **47**(8):955-961.
41. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**(7553):436-444.
42. Mamoshina P, Vieira A, Putin E, Zhavoronkov A: **Applications of Deep Learning in Biomedicine**. *Molecular pharmaceutics* 2016, **13**(5):1445-1454.
43. Alipanahi B, Delong A, Weirauch MT, Frey BJ: **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning**. *Nature biotechnology* 2015, **33**(8):831-838.
44. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model**. *Nature methods* 2015, **12**(10):931-934.
45. Quang D, Xie X: **DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences**. *Nucleic acids research* 2016.
46. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE *et al*: **Variation in transcription factor binding among humans**. *Science* 2010, **328**(5975):232-235.

47. Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L *et al*: **Effects of sequence variation on differential allelic transcription factor occupancy and gene expression**. *Genome research* 2012, **22**(5):860-869.
48. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI *et al*: **Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription**. *Science* 2013, **342**(6159):744-747.
49. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N *et al*: **AlleleSeq: analysis of allele-specific expression and binding in a network framework**. *Molecular systems biology* 2011, **7**:522.
50. Younesy H, Möller T, Heravi-Moussavi A, Cheng JB, Costello JF, Lorincz MC, Karimi MM, Jones SJM: **ALEA: a toolbox for allele-specific epigenomics analysis**. *Bioinformatics* 2014, **30**(8):1172-1174.
51. Waszak SM, Kilpinen H, Gschwind AR, Orioli A, Raghav SK, Witwicki RM, Migliavacca E, Yurovsky A, Lappalainen T, Hernandez N *et al*: **Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data**. *Bioinformatics* 2014, **30**(2):165-171.
52. Shi W, Fornes O, Mathelier A, Wasserman WW: **Evaluating the impact of single nucleotide variants on transcription factor binding**. *Nucleic acids research* 2016:gkw691.
53. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK: **Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data**. *Bioinformatics* 2009, **25**(24):3207-3212.
54. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK: **Effect of natural genetic variation on enhancer selection and function**. *Nature* 2013, **503**(7477):487-492.
55. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV *et al*: **Extensive variation in chromatin states across humans**. *Science* 2013, **342**(6159):750-752.
56. Karczewski KJ, Tatonetti NP, Landt SG, Yang X, Slifer T, Altman RB, Snyder M: **Cooperative transcription factor associations discovered using regulatory variation**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(32):13353-13358.
57. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data**. *Genome research* 2011, **21**(3):447-455.
58. Li MJ, Yan B, Sham PC, Wang J: **Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression**. *Briefings in bioinformatics* 2014.
59. Chen CC, Xiao S, Xie D, Cao X, Song CX, Wang T, He C, Zhong S: **Understanding variation in transcription factor binding by modeling transcription factor genome-epigenome interactions**. *PLoS computational biology* 2013, **9**(12):e1003367.
60. Serandour AA, Avner S, Percevault F, Demay F, Bizot M, Lucchetti-Miganeh C, Barloy-Hubler F, Brown M, Lupien M, Metivier R *et al*: **Epigenetic switch involved in**

activation of pioneer factor FOXA1-dependent enhancers. *Genome research* 2011, **21**(4):555-565.
61. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R: **Absence of a simple code: how transcription factors read the genome**. *Trends Biochem Sci* 2014, **39**(9):381-399.
62. Spivakov M, Akhtar J, Kheradpour P, Beal K, Girardot C, Koscielny G, Herrero J, Kellis M, Furlong EE, Birney E: **Analysis of variation at transcription factor binding sites in Drosophila and humans**. *Genome biology* 2012, **13**(9):R49.
63. Spivakov M: **Spurious transcription factor binding: non-functional or genetically redundant?** *BioEssays : news and reviews in molecular, cellular and developmental biology* 2014, **36**(8):798-806.
64. Dermitzakis ET, Clark AG: **Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover**. *Molecular biology and evolution* 2002, **19**(7):1114-1121.
65. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S *et al*: **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding**. *Science* 2010, **328**(5981):1036-1040.
66. Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, Myers RM, Weng Z: **Functional analysis of transcription factor binding sites in human promoters**. *Genome biology* 2012, **13**(9):R50.
67. Handstad T, Rye MB, Drablos F, Saetrom P: **A ChIP-Seq benchmark shows that sequence conservation mainly improves detection of strong transcription factor binding sites**. *PloS one* 2011, **6**(4):e18430.
68. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A *et al*: **Integrative annotation of variants from 1092 humans: application to cancer genomics**. *Science* 2013, **342**(6154):1235587.
69. Andersen MC, Engstrom PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, Odeberg J: **In silico detection of sequence variations modifying transcriptional regulation**. *PLoS computational biology* 2008, **4**(1):e5.
70. Macintyre G, Bailey J, Haviv I, Kowalczyk A: **is-rSNP: a novel technique for in silico regulatory SNP detection**. *Bioinformatics* 2010, **26**(18):i524-530.
71. Teng M, Ichikawa S, Padgett LR, Wang Y, Mort M, Cooper DN, Koller DL, Foroud T, Edenberg HJ, Econs MJ *et al*: **regSNPs: a strategy for prioritizing regulatory single nucleotide substitutions**. *Bioinformatics* 2012, **28**(14):1879-1886.
72. Manke T, Heinig M, Vingron M: **Quantifying the effect of sequence variation on regulatory interactions**. *Human mutation* 2010, **31**(4):477-483.
73. Wang J, Batmanov K: **BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations**. *Nucleic acids research* 2015, **43**(21):e147.
74. Fu Y, Liu Z, Lou S, Bedford J, Mu X, Yip KY, Khurana E, Gerstein M: **FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer**. *Genome biology* 2014, **15**(10):480.
75. Levo M, Segal E: **In pursuit of design principles of regulatory sequences**. *Nature reviews Genetics* 2014, **15**(7):453-468.

76. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications**. *Cell research* 2011, **21**(3):381-395.
77. Bai L, Morozov AV: **Gene regulation by nucleosome positioning**. *Trends in genetics : TIG* 2010, **26**(11):476-483.
78. Cui K, Zhao K: **Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq**. *Methods in molecular biology* 2012, **833**:413-419.
79. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D *et al*: **Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)**. *Genome research* 2006, **16**(1):123-131.
80. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD: **FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin**. *Genome research* 2007, **17**(6):877-885.
81. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M: **Mapping accessible chromatin regions using Sono-Seq**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(35):14926-14931.
82. Schubeler D: **Function and information content of DNA methylation**. *Nature* 2015, **517**(7534):321-326.
83. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES *et al*: **A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping**. *Cell* 2014, **159**(7):1665-1680.
84. Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, Zhang MQ, Snyder MP: **Genome-wide map of regulatory interactions in the human genome**. *Genome research* 2014, **24**(12):1905-1917.
85. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD *et al*: **In vivo enhancer analysis of human conserved non-coding sequences**. *Nature* 2006, **444**(7118):499-502.
86. Tillo D, Hughes TR: **G+C content dominates intrinsic nucleosome occupancy**. *BMC bioinformatics* 2009, **10**:442.
87. Sharif J, Endo TA, Toyoda T, Koseki H: **Divergence of CpG island promoters: a consequence or cause of evolution?** *Development, growth & differentiation* 2010, **52**(6):545-554.
88. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D *et al*: **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity**. *Genome research* 2011, **21**(10):1757-1767.
89. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position**. *Nature methods* 2013, **10**(12):1213-1218.
90. Dowell RD: **Transcription factor binding variation in the evolution of gene regulation**. *Trends in genetics : TIG* 2010, **26**(11):468-475.

91. Yanez-Cuna JO, Arnold CD, Stampfel G, Boryn LM, Gerlach D, Rath M, Stark A: **Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features**. *Genome research* 2014, **24**(7):1147-1156.
92. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: **Genome-wide quantitative enhancer activity maps identified by STARR-seq**. *Science* 2013, **339**(6123):1074-1077.
93. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Lassmann T, Itoh M *et al*: **A promoter-level mammalian expression atlas**. *Nature* 2014, **507**(7493):462-470.
94. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T *et al*: **An atlas of active enhancers across human cell types and tissues**. *Nature* 2014, **507**(7493):455-461.
95. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T *et al*: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(26):15776-15781.
96. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E *et al*: **Integrative annotation of chromatin elements from ENCODE data**. *Nucleic acids research* 2013, **41**(2):827-841.
97. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA: **High-throughput functional testing of ENCODE segmentation predictions**. *Genome research* 2014, **24**(10):1595-1602.
98. Kleftogiannis D, Kalnis P, Bajic VB: **DEEP: a general computational framework for predicting enhancers**. *Nucleic acids research* 2015, **43**(1):e6.
99. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation**. *Science* 2002, **295**(5558):1306-1311.
100. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U *et al*: **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions**. *Nature genetics* 2006, **38**(11):1341-1347.
101. Dostie J, Dekker J: **Mapping networks of physical interactions between genomic elements using 5C technology**. *Nature protocols* 2007, **2**(4):988-1002.
102. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al*: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**(5950):289-293.
103. Fullwood MJ, Ruan Y: **ChIP-based methods for the identification of long-range chromatin interactions**. *Journal of cellular biochemistry* 2009, **107**(1):30-39.
104. van Arensbergen J, van Steensel B, Bussemaker HJ: **In search of the determinants of enhancer-promoter interaction specificity**. *Trends in cell biology* 2014, **24**(11):695-702.
105. Whalen S, Truty RM, Pollard KS: **Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin**. *Nature genetics* 2016, **48**(5):488-496.

106. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M: **Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay**. *Genome research* 2013, **23**(5):800-811.
107. Maricque BB, Dougherty JD, Cohen BA: **A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells**. *Nucleic acids research* 2017, **45**(4):e16.
108. Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, Tewhey R, Isakova A, Deplancke B, Bernstein BE *et al*: **Systematic dissection of genomic features determining transcription factor binding and enhancer function**. *Proceedings of the National Academy of Sciences of the United States of America* 2017, **114**(7):E1291-E1300.
109. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM *et al*: **Massively parallel functional dissection of mammalian enhancers in vivo**. *Nature biotechnology* 2012, **30**(3):265-270.
110. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, Kellis M: **Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions**. *Nature biotechnology* 2016.
111. Cong L, Zhang F: **Genome engineering using CRISPR-Cas9 system**. *Methods in molecular biology* 2015, **1239**:197-217.
112. Han Y, Slivano OJ, Christie CK, Cheng AW, Miano JM: **CRISPR-Cas9 genome editing of a single regulatory element nearly abolishes target gene expression in mice--brief report**. *Arteriosclerosis, thrombosis, and vascular biology* 2015, **35**(2):312-315.
113. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, Banerjee B, Syed T, Emons BJ, Gifford DK, Sherwood RI: **High-throughput mapping of regulatory DNA**. *Nature biotechnology* 2016, **34**(2):167-174.
114. Lappalainen T: **Functional genomics bridges the gap between quantitative genetics and molecular biology**. *Genome research* 2015, **25**(10):1427-1431.
115. Wen X, Luca F, Pique-Regi R: **Cross-population joint analysis of eQTLs: fine mapping and functional annotation**. *PLoS genetics* 2015, **11**(4):e1005176.
116. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE *et al*: **DNase I sensitivity QTLs are a major determinant of human expression variation**. *Nature* 2012, **482**(7385):390-394.
117. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y: **Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels**. *PLoS genetics* 2014, **10**(9):e1004663.
118. Hulse AM, Cai JJ: **Genetic variants contribute to gene expression variability in humans**. *Genetics* 2013, **193**(1):95-108.
119. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S *et al*: **Genome-wide associations of gene expression variation in humans**. *PLoS genetics* 2005, **1**(6):e78.
120. Sul JH, Raj T, de Jong S, de Bakker PI, Raychaudhuri S, Ophoff RA, Stranger BE, Eskin E, Han B: **Accurate and fast multiple-testing correction in eQTL studies**. *American journal of human genetics* 2015, **96**(6):857-868.

121. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al*: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**(7468):506-511.
122. Battle A, Montgomery SB: **Determining causality and consequence of expression quantitative trait loci**. *Human genetics* 2014, **133**(6):727-735.
123. Gibson G: **Rare and common variants: twenty arguments**. *Nature reviews Genetics* 2012, **13**(2):135-145.
124. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium GT, Nicolae DL *et al*: **A gene-based association method for mapping traits using reference transcriptome data**. *Nature genetics* 2015, **47**(9):1091-1098.
125. Manor O, Segal E: **Robust prediction of expression differences among human individuals using only genotype information**. *PLoS genetics* 2013, **9**(3):e1003396.
126. Vervier K, Michaelson JJ: **SLINGER: large-scale learning for predicting gene expression**. *Scientific reports* 2016, **6**:39360.
127. Wen X, Lee Y, Luca F, Pique-Regi R: **Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors**. *American journal of human genetics* 2016, **98**(6):1114-1129.
128. Waszak SM, Delaneau O, Gschwind AR, Kilpinen H, Raghav SK, Witwicki RM, Orioli A, Wiederkehr M, Panousis NI, Yurovsky A *et al*: **Population Variation and Genetic Control of Modular Chromatin Architecture in Humans**. *Cell* 2015, **162**(5):1039-1050.
129. Gaffney DJ: **Global properties and functional complexity of human gene regulatory variation**. *PLoS genetics* 2013, **9**(5):e1003501.
130. Lecerf L, Kavo A, Ruiz-Ferrer M, Baral V, Watanabe Y, Chaoui A, Pingault V, Borrego S, Bondurand N: **An impairment of long distance SOX10 regulatory elements underlies isolated Hirschsprung disease**. *Human mutation* 2014, **35**(3):303-307.
131. Smemo S, Campos LC, Moskowitz IP, Krieger JE, Pereira AC, Nobrega MA: **Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease**. *Human molecular genetics* 2012, **21**(14):3255-3263.
132. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M: **A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals**. *Nature communications* 2016, **7**:11101.
133. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.
134. Harvey CT, Moyerbrailean GA, Davis GO, Wen X, Luca F, Pique-Regi R: **QuASAR: quantitative allele-specific analysis of reads**. *Bioinformatics* 2015, **31**(8):1235-1242.
135. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S *et al*: **Annotation of functional variation in personal genomes using RegulomeDB**. *Genome research* 2012, **22**(9):1790-1797.
136. Herrmann C, Van de Sande B, Potier D, Aerts S: **i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules**. *Nucleic acids research* 2012, **40**(15):e114.

137. Karczewski KJ, Dudley JT, Kukurba KR, Chen R, Butte AJ, Montgomery SB, Snyder M: **Systematic functional regulatory assessment of disease-associated variants**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(23):9607-9612.
138. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J *et al*: **Defining functional DNA elements in the human genome**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(17):6131-6138.
139. Bailey SD, Virtanen C, Haibe-Kains B, Lupien M: **ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments**. *Bioinformatics* 2015, **31**(18):3057-3059.
140. Zuo C, Shin S, Keles S: **atSNP: transcription factor binding affinity testing for regulatory SNP detection**. *Bioinformatics* 2015, **31**(20):3353-3355.
141. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G *et al*: **Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays**. *Science* 2010, **327**(5961):78-81.
142. Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, Qiu R, Lee C, Shendure J: **The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line**. *Nature* 2013, **500**(7461):207-211.
143. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M *et al*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nature genetics* 2011, **43**(5):491-498.
144. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
145. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P *et al*: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. *Genome research* 2012, **22**(9):1813-1831.
146. Wilbanks EG, Facciotti MT: **Evaluation of algorithm performance in ChIP-seq peak detection**. *PloS one* 2010, **5**(7):e11471.
147. Worsley Hunt R, Mathelier A, Del Peso L, Wasserman WW: **Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment**. *BMC genomics* 2014, **15**:472.
148. Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, Stamatoyannopoulos JA: **Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo**. *Nature genetics* 2015.
149. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H *et al*: **JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles**. *Nucleic acids research* 2014, **42**(Database issue):D142-147.
150. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**(11):1422-1423.

151. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C *et al*: **RSAT 2015: Regulatory Sequence Analysis Tools**. *Nucleic acids research* 2015, **43**(W1):W50-56.
152. Breiman L: **Random forests**. *Machine learning* 2001, **45**(1):5-32.
153. Kuhn M: **caret: Classification and Regression Training**. 2015.
154. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR: **A balanced iterative random forest for gene selection from microarray data**. *BMC bioinformatics* 2013, **14**:261.
155. Bekkar M, Alitouche TA: **Imbalanced Data Learning Approaches Review**. *International Journal of Data Mining & Knowledge Management Process* 2013, **3**(4):15-33.
156. Melton C, Reuter JA, Spacek DV, Snyder M: **Recurrent somatic mutations in regulatory regions of human cancer genomes**. *Nature genetics* 2015, **47**(7):710-716.
157. Mathelier A, Lefebvre C, Zhang AW, Arenillas DJ, Ding J, Wasserman WW, Shah SP: **Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas**. *Genome biology* 2015, **16**:84.
158. Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC *et al*: **Cooperativity and rapid evolution of cobound transcription factors in closely related mammals**. *Cell* 2013, **154**(3):530-540.
159. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic acids research* 2006, **34**(Database issue):D535-539.
160. Cavalli M, Pan G, Nord H, Wallen Arzt E, Wallerman O, Wadelius C: **Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals**. *Genomics* 2016.
161. Nakahashi H, Kwon KR, Resch W, Vian L, Dose M, Stavreva D, Hakim O, Pruett N, Nelson S, Yamane A *et al*: **A genome-wide map of CTCF multivalency redefines the CTCF code**. *Cell reports* 2013, **3**(5):1678-1689.
162. Maerkl SJ, Quake SR: **A systems approach to measuring the binding energy landscapes of transcription factors**. *Science* 2007, **315**(5809):233-237.
163. Nutiu R, Friedman RC, Luo S, Khrebtukova I, Silva D, Li R, Zhang L, Schroth GP, Burge CB: **Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument**. *Nature biotechnology* 2011, **29**(7):659-664.
164. Levo M, Zalckvar E, Sharon E, Dantas Machado AC, Kalma Y, Lotam-Pompan M, Weinberger A, Yakhini Z, Rohs R, Segal E: **Unraveling determinants of transcription factor binding outside the core binding site**. *Genome research* 2015, **25**(7):1018-1029.
165. Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R: **DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale**. *Nucleic acids research* 2013, **41**(Web Server issue):W56-62.
166. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordan R, Rohs R: **Quantitative modeling of transcription factor binding specificities using DNA shape**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(15):4654-4659.
167. Cavalli M, Pan G, Nord H, Wallerman O, Wallen Arzt E, Berggren O, Elvers I, Eloranta ML, Ronnblom L, Lindblad Toh K *et al*: **Allele-specific transcription factor binding to**

common and rare variants associated with disease and gene expression. *Human genetics* 2016, **135**(5):485-497.
168. Ding Z, Ni Y, Timmer SW, Lee BK, Battenhouse A, Louzada S, Yang F, Dunham I, Crawford GE, Lieb JD *et al*: **Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association**. *PLoS genetics* 2014, **10**(11):e1004798.
169. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome biology* 2010, **11**(10):R106.
170. van de Geijn B, McVicker G, Gilad Y, Pritchard JK: **WASP: allele-specific software for robust molecular quantitative trait locus discovery**. *Nature methods* 2015, **12**(11):1061-1063.
171. Yee TW: **The VGAM package for categorical data analysis**. *Journal of statistical software* 2010, **32**(10):1-34.
172. Team RC: **R: A language and environment for statistical computing**. 2013.
173. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM: **A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data**. *Genome research* 2011, **21**(10):1728-1737.
174. Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, Watanabe C, Zhang Z: **MBASED: allele-specific expression detection in cancer tissues and cell lines**. *Genome biology* 2014, **15**(8):405.
175. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(23):9362-9367.
176. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA *et al*: **Genetic and epigenetic fine mapping of causal autoimmune disease variants**. *Nature* 2014.
177. Dror I, Rohs R, Mandel-Gutfreund Y: **How motif environment influences transcription factor search dynamics: Finding a needle in a haystack**. *BioEssays : news and reviews in molecular, cellular and developmental biology* 2016.
178. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y: **The functional consequences of variation in transcription factor binding**. *PLoS genetics* 2014, **10**(3):e1004226.
179. Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, Greenside P, Srivas R, Phanstiel DH, Pekowska A *et al*: **Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions**. *Cell* 2015, **162**(5):1051-1065.
180. Patro R, Mount SM, Kingsford C: **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms**. *Nature biotechnology* 2014, **32**(5):462-464.
181. Stegle O, Parts L, Piipari M, Winn J, Durbin R: **Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses**. *Nature protocols* 2012, **7**(3):500-507.
182. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK *et al*: **An expansive human regulatory lexicon encoded in transcription factor footprints**. *Nature* 2012, **489**(7414):83-90.

183. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent**. *Journal of statistical software* 2010, **33**(1):1-22.
184. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L *et al*: **The BioGRID interaction database: 2015 update**. *Nucleic acids research* 2015, **43**(Database issue):D470-478.
185. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets**. *GigaScience* 2015, **4**:7.
186. Hiller M, Agarwal S, Notwell JH, Parikh R, Guturu H, Wenger AM, Bejerano G: **Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish**. *Nucleic acids research* 2013, **41**(15):e151.
187. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M *et al*: **Patterns of cis regulatory variation in diverse human populations**. *PLoS genetics* 2012, **8**(4):e1002639.
188. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis**. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2014, **47**:11 12 11-34.
189. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**(15):2156-2158.
190. Wickham H: **ggplot2: elegant graphics for data analysis**: Springer Science & Business Media; 2009.
191. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**(7422):56-65.
192. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U: **Predicting cell-type-specific gene expression from regions of open chromatin**. *Genome research* 2012, **22**(9):1711-1722.
193. Ong CT, Corces VG: **CTCF: an architectural protein bridging genome topology and function**. *Nature reviews Genetics* 2014, **15**(4):234-246.
194. Lotem J, Levanon D, Negreanu V, Bauer O, Hantisteanu S, Dicken J, Groner Y: **Runx3 at the interface of immunity, inflammation and cancer**. *Biochimica et biophysica acta* 2015, **1855**(2):131-143.
195. Hagman J, Ramirez J, Lukin K: **B lymphocyte lineage specification, commitment and epigenetic control of transcription by early B cell factor 1**. *Current topics in microbiology and immunology* 2012, **356**:17-38.
196. Iwafuchi-Doi M, Zaret KS: **Pioneer transcription factors in cell reprogramming**. *Genes & development* 2014, **28**(24):2679-2692.
197. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Zappala Z, Strober BJ, Scott AJ, Ganna A *et al*: **The impact of rare variation on gene expression across tissues**. 2016.
198. Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA: **Widespread site-dependent buffering of human regulatory polymorphism**. *PLoS genetics* 2012, **8**(3):e1002599.

199. Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y: **A widespread role of the motif environment in transcription factor binding across diverse protein families**. *Genome research* 2015, **25**(9):1268-1280.
200. Rajkumar AS, Denervaud N, Maerkl SJ: **Mapping the fine structure of a eukaryotic promoter input-output function**. *Nature genetics* 2013, **45**(10):1207-1215.
201. Mathelier A, Xin B, Chiu TP, Yang L, Rohs R, Wasserman WW: **DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo**. *Cell systems* 2016, **3**(3):278-286 e274.
202. Chen CY, Shi W, Balaton BP, Matthews AM, Li Y, Arenillas DJ, Mathelier A, Itoh M, Kawaji H, Lassmann T *et al*: **YY1 binding association with sex-biased transcription revealed through X-linked transcript levels and allelic binding analyses**. *Scientific reports* 2016, **6**:37324.
203. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I: **Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers**. *Genome research* 2010, **20**(5):565-577.
204. Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A: **Assessing computational methods for transcription factor target gene identification based on ChIP-seq data**. *PLoS computational biology* 2013, **9**(11):e1003342.
205. Chen X, Jung JG, Shajahan-Haq AN, Clarke R, Shih Ie M, Wang Y, Magnani L, Wang TL, Xuan J: **ChIP-BIT: Bayesian inference of target genes using a novel joint probabilistic model of ChIP-seq profiles**. *Nucleic acids research* 2016, **44**(7):e65.
206. Caliskan M, Cusanovich DA, Ober C, Gilad Y: **The effects of EBV transformation on gene expression levels and methylation profiles**. *Human molecular genetics* 2011, **20**(8):1643-1652.
207. Sie L, Loong S, Tan EK: **Utility of lymphoblastoid cell lines**. *Journal of neuroscience research* 2009, **87**(9):1953-1959.
208. Consortium GT: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans**. *Science* 2015, **348**(6235):648-660.
209. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, Watt S, Yan Y, Kundu K, Ecker S *et al*: **Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells**. *Cell* 2016, **167**(5):1398-1414 e1324.
210. Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL: **Epigenetic priors for identifying active transcription factor binding sites**. *Bioinformatics* 2012, **28**(1):56-62.
211. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J *et al*: **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015, **518**(7539):317-330.
212. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D *et al*: **Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells**. *Nature* 2013, **498**(7453):236-240.
213. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzetinova T *et al*: **Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data**. *Lancet* 2015, **385**(9975):1305-1314.

214. Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, Hijazi H, Alshammari M, Aldahmesh MA, Salih MA *et al*: **Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families**. *Cell reports* 2015, **10**(2):148-161.
215. Tarailo-Graovac M, Shyr C, Ross CJ, Horvath GA, Salvarinova R, Ye XC, Zhang LH, Bhavsar AP, Lee JJ, Drogemoller BI *et al*: **Exome Sequencing and the Management of Neurometabolic Disorders**. *The New England journal of medicine* 2016, **374**(23):2246-2255.
216. Campeau PM, Foulkes WD, Tischkowitz MD: **Hereditary breast cancer: new genetic developments, new therapeutic avenues**. *Human genetics* 2008, **124**(1):31-42.
217. Guturu H, Chinchali S, Clarke SL, Bejerano G: **Erosion of Conserved Binding Sites in Personal Genomes Points to Medical Histories**. *PLoS computational biology* 2016, **12**(2):e1004711.
218. Roy S, Guler R, Parihar SP, Schmeier S, Kaczkowski B, Nishimura H, Shin JW, Negishi Y, Ozturk M, Hurdayal R *et al*: **Batf2/Irf1 induces inflammatory responses in classically activated macrophages, lipopolysaccharides, and mycobacterial infection**. *Journal of immunology* 2015, **194**(12):6035-6044.
219. Cherasse Y, Maurin AC, Chaveroux C, Jousse C, Carraro V, Parry L, Deval C, Chambon C, Fafournoux P, Bruhat A: **The p300/CBP-associated factor (PCAF) is a cofactor of ATF4 for amino acid-regulated transcription of CHOP**. *Nucleic acids research* 2007, **35**(17):5954-5965.
220. Kim WH, Jang MK, Kim CH, Shin HK, Jung MH: **ATF3 inhibits PDX-1-stimulated transactivation**. *Biochemical and biophysical research communications* 2011, **414**(4):681-687.

Appendices

Appendix A Supplementary material for Chapter 2

A.1 Replicate normalization method produces highly similar sets of ASB calls

In Chapter 2, we used a direct sum approach in which we summed the read counts of each allele across the replicates, and then applied the binomial test on the derived sum of each allele. We also implemented a normalized approach regarding multiple replicates and compared the ASB calling between the two (direct sum and normalized). In the normalization approach, the read coverage at heterozygous positions is normalized between replicates following the scale factor-based procedures used in DEseq [169]. The normalized count of each site is the original count divided by the scale factor. The normalized count values are thereafter processed using the same procedure as in direct sum approach.

The normalized approach resulted in 10,121 called ASB events, while direct-sum approach called 10,711. Overall, 9,511 ASB events were called by both approaches, and on average, 92% of the called ASB events using the normalized approach overlapped with those called with the direct-sum approach across investigated TFs. While a few datasets showed greater difference, as shown in Appendix Figure A8, most samples were clustered in the lower right corner reflecting high similarity between the results.

As one would expect, we observed that the overlap ratio was anti-correlated with the scale factor of the larger replicate (Spearman correlation coefficient = -0.64; Appendix Figure A8), showing a large replicate library difference in depth (large scale factor) correlated with greater divergence

in ASB calling between the two approaches. However, the impact was modest, as there was still 85-95% overlap for the larger scale factor cases.

We explored the differences between methods, which confirmed our expectation that the normalized approach penalized those cases in which one replicate was strikingly lower than the other in terms of counts. This can be observed in Appendix Figure A9, in which we show the read coverage for the method-specific ASB calls for a TF experiment with high scale factor.

A.2 Direct sum approach is used considering the characteristics of the data

It is useful to recognize two key aspects of the ENCODE data prior to reviewing the findings. For the vast majority of TFs there are only two replicates (n=41), with only a few having three replicates (n=4). The second is a difference between standard RNA-seq and ASB identification in ChIP-seq. In standard RNA-seq (as most published methods address), each sample is prepared and processed separately. For ASB detection in ChIP-seq, two alleles are naturally controlled within the same single sample. These two aspects inform our decision about the selection of the ASB calling method.

Based on our perspective, with only two replicates for the vast majority of cases, we prefer to use the direct-sum approach. This reflects our view that the high coverage positions in a single ChIP-seq replicate are well controlled (two alleles coming from the same nuclei). We anticipate that replicate normalization will be an important issue and should be deeply considered in future ASB analysis (particularly when greater replicate numbers are available).

A.3 The sequence based classifier produces consistent predictions for lymphoblastoid cells across multiple individuals

We tested the consistency of the random forest classifier in different individuals from the same cell type (that is lymphoblastoid cell line). Briefly, we collected CTCF ChIP-seq data from multiple ENCODE samples. We trained a sequence based classifier with N-1 samples (N is the number of collected samples), and tested each model on the remaining sample. Results showed similar performance between cross validation and testing (for instance, the mean AUPRC difference is equal to 0.02 and the standard deviation is 0.05, Appendix Figure A5). These results suggested that our sequence based model could be applied across individuals using a single training data set.

A.4 Supplementary figures and tables for Chapter 2

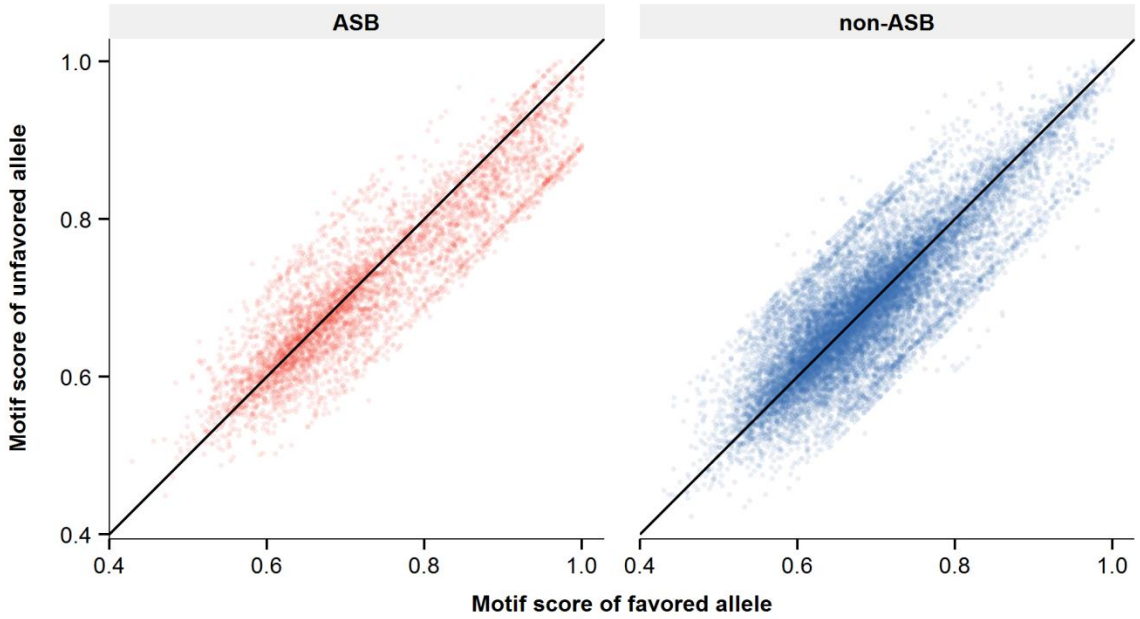


Figure A1. Comparing motif score of two alleles for all the investigated heterozygous site binding events. Each dot represents the relative motif score for favored allele (allele with higher ChIP-seq read count) and unfavored allele. ASB and non-ASB events are plotted separately. The black diagonal line indicates the cases with equal scores of two alleles. Heterozygous site binding data of all the TFs with known motif are presented together. This figure presents the entire data set partially depicted in Figure 2.1.

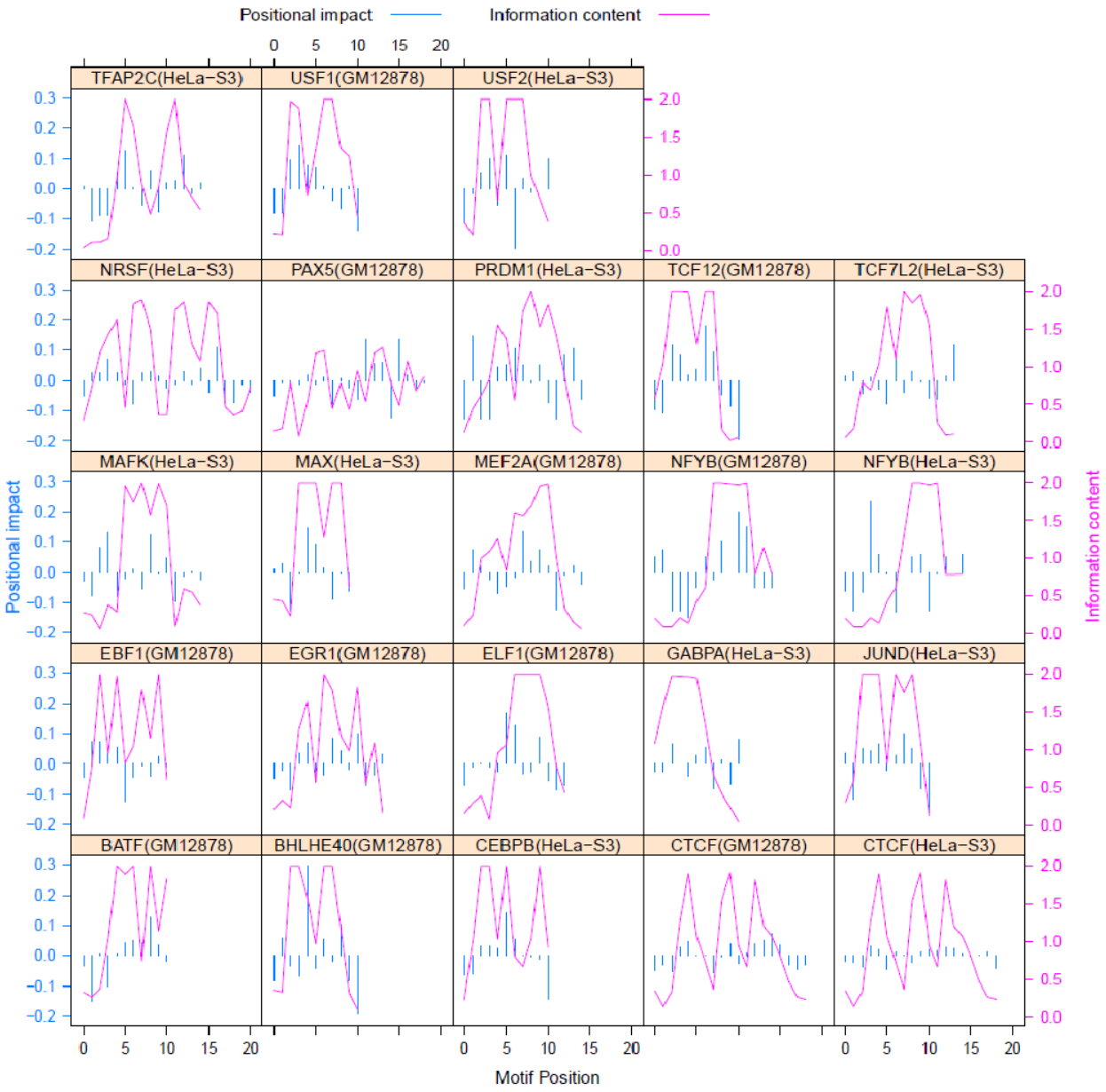


Figure A2. The positional impact and information content at each position of TF motifs.

For each TF, we plot the positional impact of each motif position derived from ASB events (red bar) and its corresponding information content (blue line). This figure presents a TF specific perspective of Figure 2.2A.

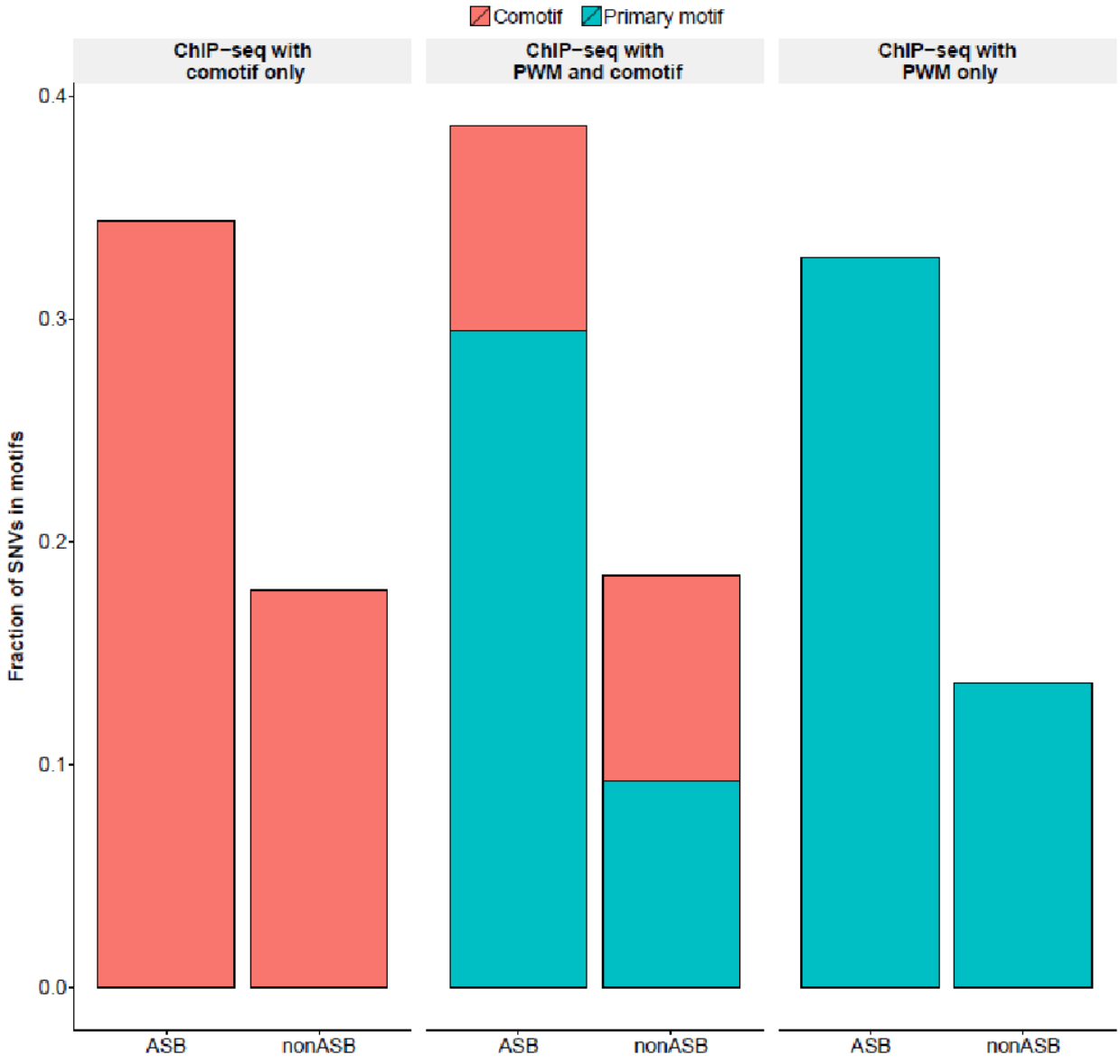


Figure A3. SNVs of ASB events are enriched in predicted TFBSs and comotifs.

The ChIP-seq experiments are divided into TFs with comotif only (left panel; TFs with no known PWM), with both PWM and comotifs (middle panel), and with known PWMs only (right panel). ASB-SNVs are significantly enriched in the predicted TFBS of comotifs (p-value = 2.2×10^{-16}, odds ratio = 2.4, Fisher's exact test), combination of comotif and primary motif (p-value = 2.2×10^{-16}, odds ratio = 2.7, Fisher's exact test) and primary motif (p-value = 2.2×10^{-16}, odds ratio = 3.0, Fisher's exact test).

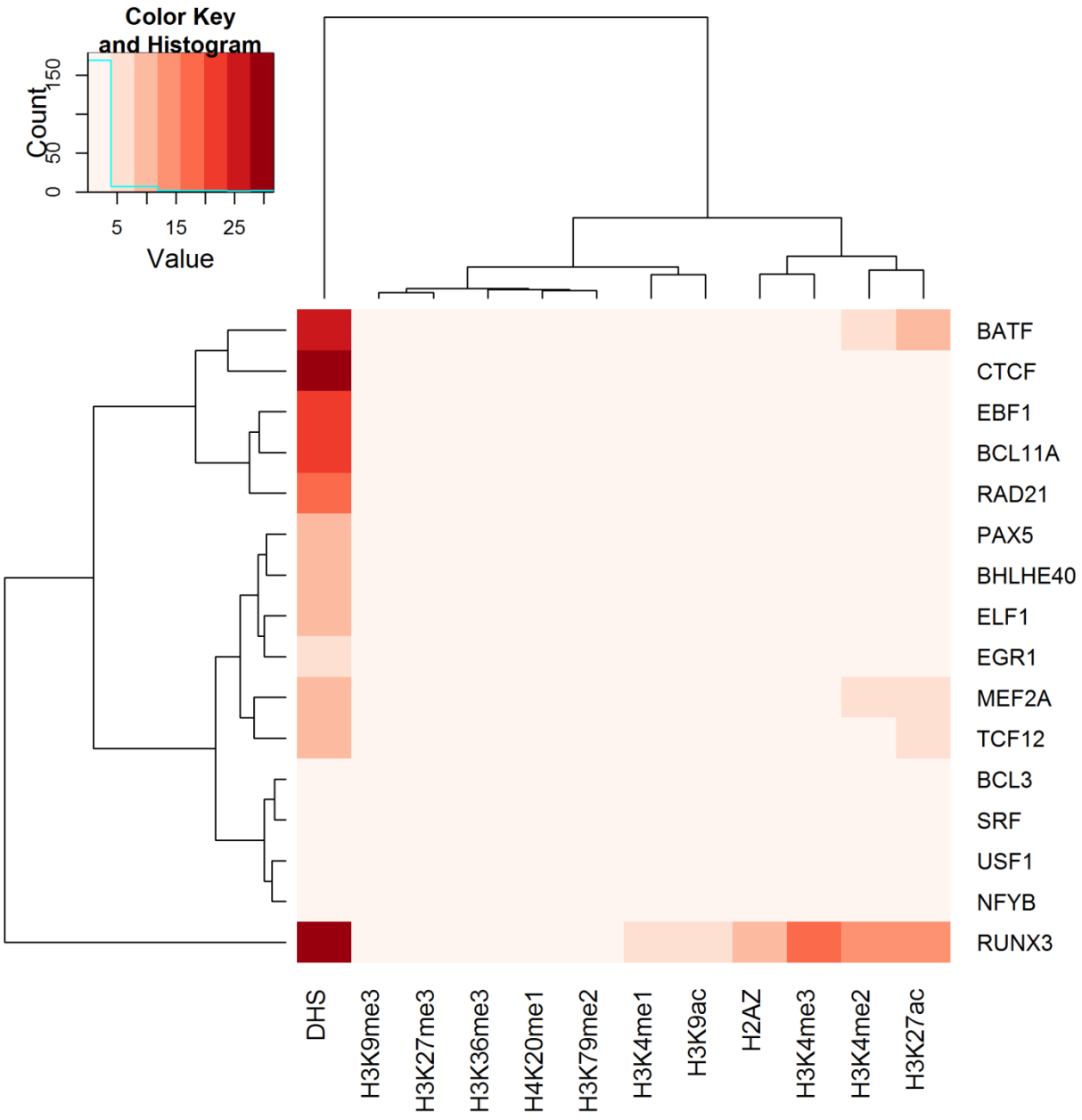


Figure A4. Allelic coordination between TFs and chromatin properties in GM12878 cell line.

The heatmap represents the -log(p-value) of Pearson correlation between allele imbalance of TF ChIP-seq reads at heterozygous site binding events and chromatin properties (DHS and histone modifications).

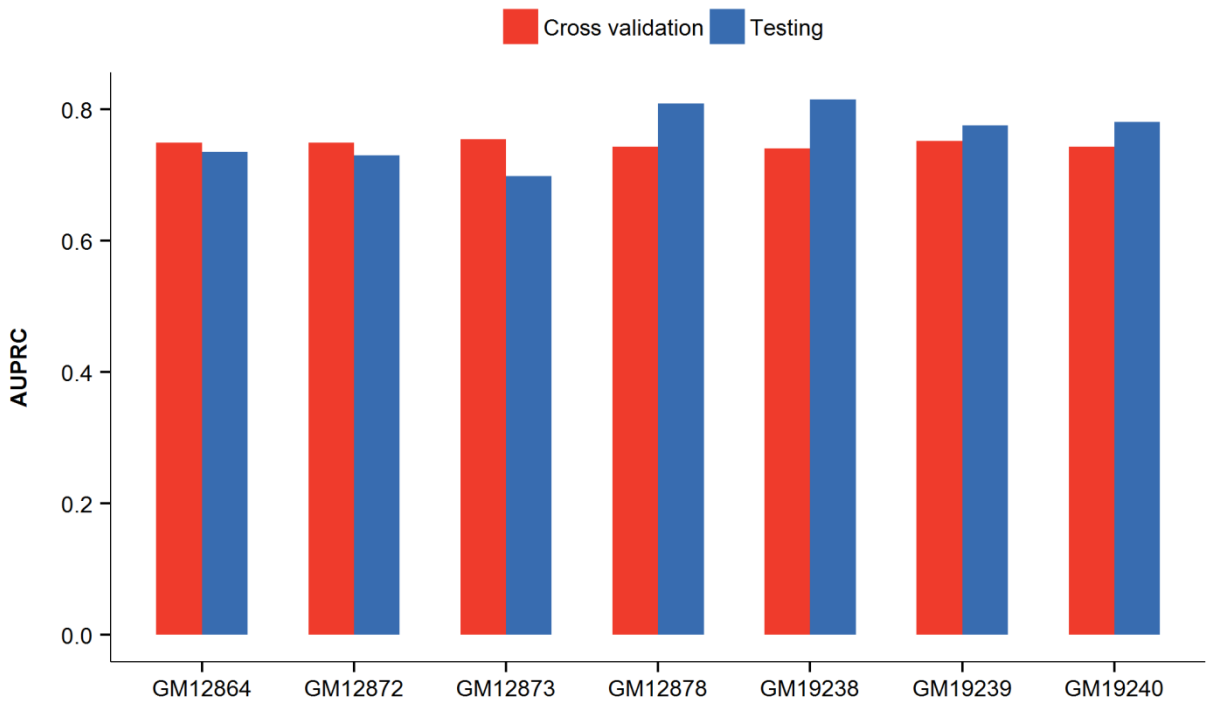


Figure A5. Testing the performance of sequence models for CTCF in seven individuals.

The figure shows the accuracy of cross validation within any six samples (red bar) and testing accuracy of the remaining individual (blue bar).

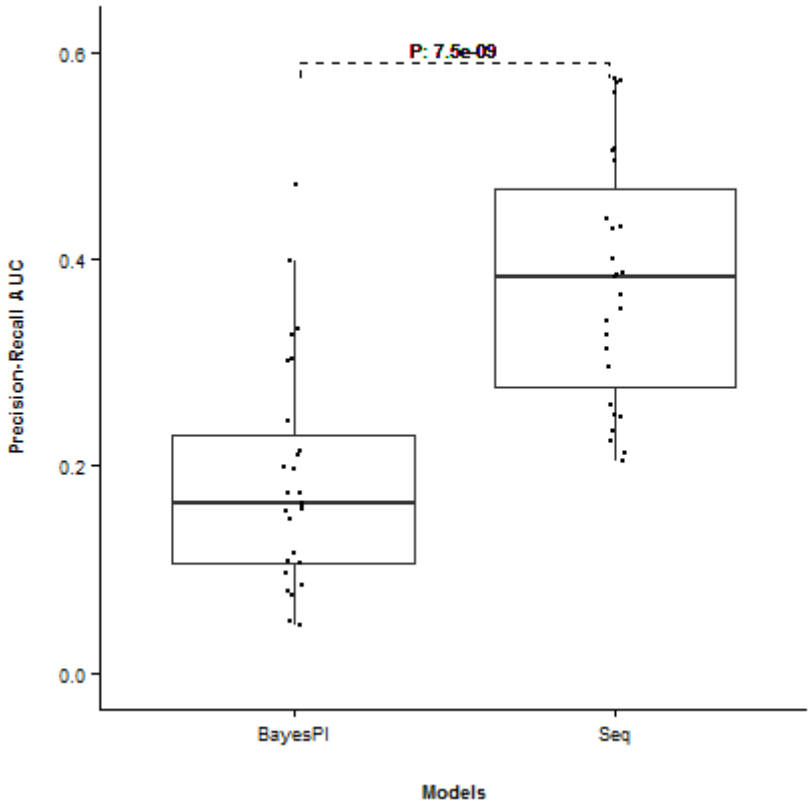


Figure A6. Compare the performance of the Seq model and BayesPI-Bar. Only 27 TFs experiments with available BayesPI-Bar models are presented in the comparison.

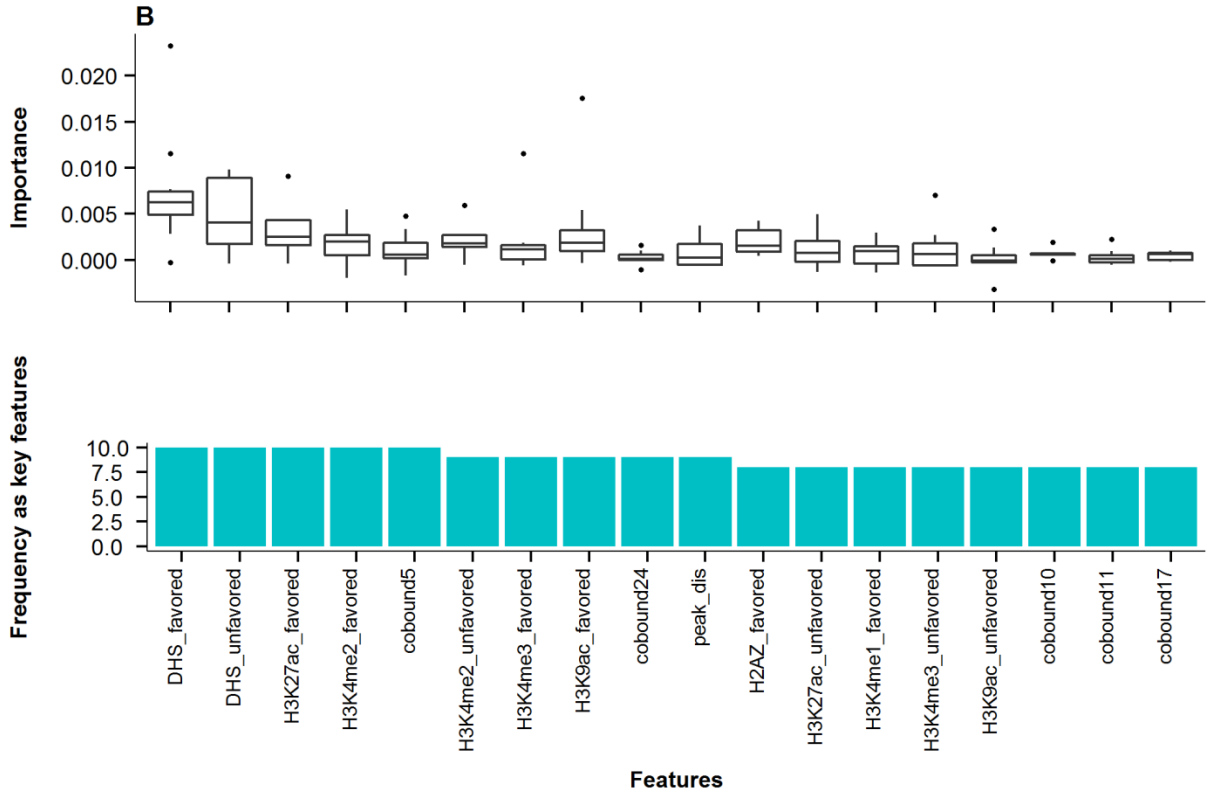


Figure A7 The most frequently selected key features in the Full models for the TFs without known motif

The suffix 'favor' (respectively 'unfavor') refers to the allele with higher (respectively lower) read counts at heterozygous sites. Details of each feature can be found in the Methods section and Appendix Table A5.

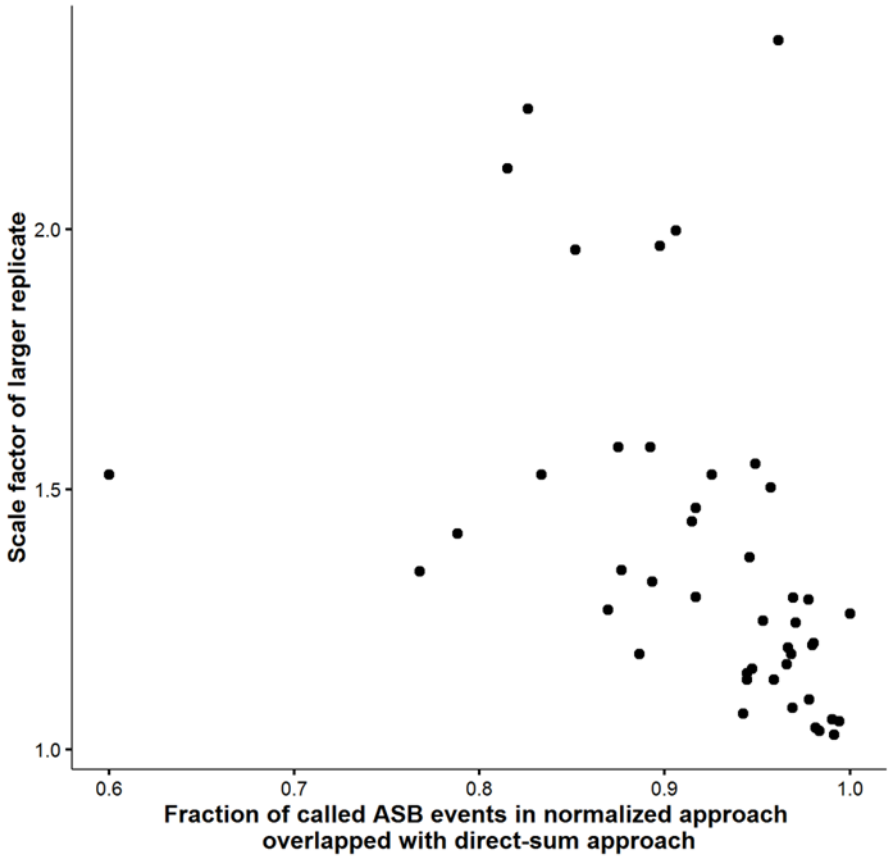


Figure A8 Scale factor and overlap ratio between the direct sum and the normalized approach.

Each point represents one TF ChIP-seq dataset.

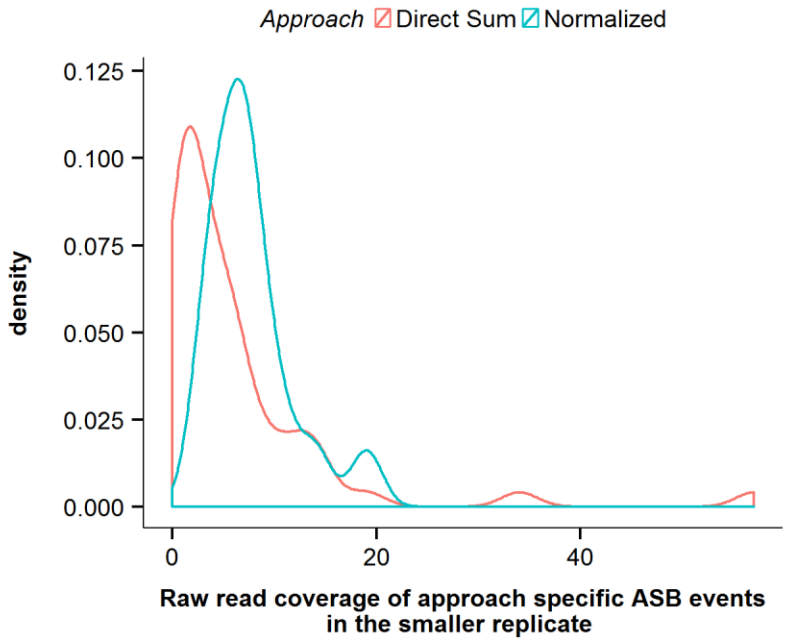


Figure A9 Read coverage distribution of approach-specific ASB events in the smaller replicate

The data of CHD2 in HeLa-S3 are shown as it has a low overlap ratio (82.6%) and a high scale factor 2.1. In this TF experiment, direct sum approach called 129 ASB events and normalized approach called 92, of which 76 events are overlapped between two approaches.

Data Type	Institution	URL
TF ChIP-seq raw reads	Haib	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/
TF ChIP-seq raw reads	Sydh	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/
TF ChIP-seq raw reads	Uta	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromChip/
TF ChIP-seq raw reads	Uw	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwTfbs/
TF narrowPeak regions	Awg	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/
DNase-Seq raw reads	Uw	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/
DNase-Seq raw reads	Duke	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/
Histone ChIP-seq raw reads	Broad	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/
Genotype data of Lymphoblastoid cell lines	Complete genomics	ftp://ftp2.completegenomics.com/vcf_files/Build37_2.0.0/
Copy number variant region	Haib	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibGenotype/ (Available for GM12878, HeLa-S3, and GM19238)

Table A1. Sources of the data used in ASB analysis.

Our analysis integrated multiple types of data, including ChIP-seq, DNase-Seq, and genotype calling data. The categories, source institution, and URL of these data were listed in the table.

Cell	TF	Peak Count	Heterozygous binding sites events	ASB
GM12878	BATF	32427	1314	201
GM12878	BCL11A	17876	678	60
GM12878	BCL3	15455	653	60
GM12878	BHLHE40	13986	661	66
GM12878	CTCF	55551	2614	396
GM12878	EBF1	33410	1445	327
GM12878	EGR1	16331	718	69
GM12878	ELF1	23008	1047	58
GM12878	MEF2A	17605	574	55
GM12878	NFYB	13295	355	77
GM12878	PAX5	19740	657	43
GM12878	RAD21	40019	1963	281
GM12878	RUNX3	67965	3304	471
GM12878	SRF	8544	164	31
GM12878	TCF12	20437	770	82
GM12878	USF1	9778	305	37
HELA-S3	TFAP2C	25452	561	84
HELA-S3	BRCA1	8114	274	47
HELA-S3	CEBPB	61004	2491	922
HELA-S3	CHD2	20500	696	129
HELA-S3	CTCF	58806	2506	1076
HELA-S3	GABPA	6761	224	72
HELA-S3	JUND	31633	641	118
HELA-S3	MAFK	14185	431	84
HELA-S3	MAX	29647	1111	132
HELA-S3	NFYB	7156	149	47
HELA-S3	NRSF	10247	372	179

Cell	TF	Peak Count	Heterozygous binding sites events	ASB
HELA-S3	P300	25854	661	160
HELA-S3	POL2	25332	1284	570
HELA-S3	PRDM1	4577	134	50
HELA-S3	RAD21	43420	1883	708
HELA-S3	RFX5	19284	664	80
HELA-S3	SMC3	39567	1860	565
HELA-S3	STAT3	13834	325	56
HELA-S3	TAF1	16100	514	83
HELA-S3	TBP	18489	466	87
HELA-S3	TCF7L2	19242	600	124
HELA-S3	USF2	12306	373	108
HELA-S3	ZNF143	7048	261	52
GM12872	CTCF	47151	2496	488
GM12873	CTCF	51005	2575	552
GM19238	CTCF	49938	2909	500
GM19239	CTCF	41085	2473	282
GM19240	CTCF	46036	2972	573
GM12864	CTCF	46798	2390	523
Total	45	1205998	51518	10765

Table A2. Processed heterozygous site binding data.

For each TF ChIP-seq experiment, we listed the number of ChIP-seq peaks (Peak count), heterozygous site binding events, and called ASB events.

Cell	TF	Comotif
GM12878	BATF	IRF1.IRF, BZIP.IRF[218]
GM12878	BHLHE40	RUNX1.RUNT
GM12878	RAD21	CTCF.ZF[184]
GM12878	RUNX3	BATF.BZIP, RUNX1.RUNT, ETS1.ETS, FLI1.ETS
HeLa-S3	CEBPB	BATF.BZIP
HeLa-S3	P300	NF-E2.BZIP, CEBP.BZIP[184], ATF3.BZIP[219, 220]
HeLa-S3	RAD21	CTCF.ZF[184]
HeLa-S3	SMC3	CTCF.ZF[184]
HeLa-S3	TCF7L2	ATF3.BZIP
Total	9	15

Table A3. Discovered comotifs from heterozygous site binding events.

HOMER motifs were considered as comotifs when their motif change correlated with TF allelic binding imbalance in heterozygous site binding events (see Materials and methods in Chapter2). The cell line, ChIP'ed TF, and correlated comotifs are provided respectively.TF-comotif pairs supported by external literature are given the corresponding references.

Cell	ASB-TF	TF	-log(p-value)	Odd ratio
GM12878	BATF	TBP	4.50	0.25
GM12878	BATF	CDP	3.94	0.49
GM12878	BATF	POL24H8	3.51	0.29
GM12878	BATF	STAT3	3.36	0.31
GM12878	CTCF	ZNF143	10.45	0.43
GM12878	CTCF	YY1	7.42	0.48
GM12878	CTCF	POU2F2	4.83	0.37
GM12878	CTCF	FOXM1	3.76	0.49
GM12878	CTCF	SMC3	3.36	0.68
GM12878	EBF1	TBP	3.78	0.42
GM12878	EBF1	NFKB	3.76	0.54
GM12878	EBF1	IKZF1	3.75	0.41
GM12878	EBF1	MXI1	3.52	0.48
GM12878	MEF2A	PML	3.83	0.28
GM12878	MEF2A	MAX	3.51	0.00
GM12878	MEF2A	BCLAF1	3.46	0.13
GM12878	RAD21	ZNF143	9.04	0.42
GM12878	RAD21	RUNX3	5.61	0.47
GM12878	RAD21	PAX5	4.69	0.37
GM12878	RAD21	YY1	4.08	0.57
GM12878	RAD21	FOXM1	3.43	0.51
GM12878	RUNX3	POL2	12.31	0.34
GM12878	RUNX3	ZNF143	9.73	0.29
GM12878	RUNX3	POL24H8	9.29	0.37
GM12878	RUNX3	ELF1	8.16	0.42
GM12878	RUNX3	SMC3	7.86	0.37
GM12878	RUNX3	YY1	7.71	0.48
GM12878	RUNX3	CTCF	7.55	0.36

Cell	ASB-TF	TF	-log(p-value)	Odd ratio
GM12878	RUNX3	MAZ	6.71	0.46
GM12878	RUNX3	PML	6.08	0.49
GM12878	RUNX3	ELK1	5.65	0.28
GM12878	RUNX3	EGR1	5.34	0.42
GM12878	RUNX3	GR	5.34	0.42
GM12878	RUNX3	RAD21	5.12	0.49
GM12878	RUNX3	SIN3A	5.12	0.39
GM12878	RUNX3	CMYC	5.01	0.17
GM12878	RUNX3	ZEB1	5.01	0.17
GM12878	RUNX3	MAX	4.82	0.47
GM12878	RUNX3	GABP	4.61	0.31
GM12878	RUNX3	TAF1	4.47	0.49
GM12878	RUNX3	CHD2	4.45	0.54
GM12878	RUNX3	BATF	4.05	1.52
GM12878	RUNX3	TCF3	3.90	0.56
GM12878	RUNX3	TBP	3.80	0.56
GM12878	RUNX3	SRF	3.62	0.41
GM12878	RUNX3	MXI1	3.56	0.58
GM12878	USF1	SIN3A	4.26	0.06
HeLa-S3	CEBPB	P300	16.61	2.07
HeLa-S3	CEBPB	STAT3	6.52	1.68
HeLa-S3	CEBPB	CTCF	4.56	0.50
HeLa-S3	CEBPB	CJUN	3.58	1.41
HeLa-S3	CEBPB	JUND	3.56	1.38
HeLa-S3	CEBPB	TCF7L2	3.40	1.45
HeLa-S3	CTCF	SMC3	7.44	1.59
HeLa-S3	CTCF	RAD21	6.59	1.58
HeLa-S3	MAX	CMYC	6.77	3.31

Cell	ASB-TF	TF	-log(p-value)	Odd ratio
HeLa-S3	MAX	USF2	4.59	2.27
HeLa-S3	P300	ELK4	3.20	3.10
HeLa-S3	POL2	TAF1	10.46	2.14
HeLa-S3	POL2	HCFC1	8.65	1.98
HeLa-S3	POL2	TBP	6.50	1.80
HeLa-S3	POL2	E2F4	5.36	2.43
HeLa-S3	POL2	GCN5	4.76	2.95
HeLa-S3	POL2	P300	3.89	0.50
HeLa-S3	POL2	CJUN	3.72	0.52
HeLa-S3	POL2	NRF1	3.24	2.11
HeLa-S3	RAD21	CTCF	16.29	2.58
HeLa-S3	RAD21	CJUN	10.37	0.35
HeLa-S3	RAD21	P300	9.96	0.38
HeLa-S3	RAD21	CFOS	9.38	0.28
HeLa-S3	RAD21	COREST	7.49	0.48
HeLa-S3	RAD21	REST	7.49	0.48
HeLa-S3	RAD21	CHD2	7.42	0.46
HeLa-S3	RAD21	TCF7L2	7.15	0.44
HeLa-S3	RAD21	SMC3	6.33	1.94
HeLa-S3	RAD21	STAT3	5.60	0.47
HeLa-S3	RAD21	TBP	4.84	0.50
HeLa-S3	RAD21	POL2	4.11	0.49
HeLa-S3	RAD21	TAF1	4.01	0.44
HeLa-S3	RAD21	BAF155	3.97	0.47
HeLa-S3	RAD21	MXI1	3.97	0.47
HeLa-S3	RAD21	MAX	3.78	0.65
HeLa-S3	RAD21	CEBPB	3.66	0.66
HeLa-S3	RAD21	JUND	3.22	0.65

Cell	ASB-TF	TF	-log(p-value)	Odd ratio
HeLa-S3	RFX5	SMC3	3.96	0.38
HeLa-S3	SMC3	CTCF	17.47	3.03
HeLa-S3	SMC3	RAD21	7.77	2.36
HeLa-S3	SMC3	CJUN	6.45	0.41
HeLa-S3	SMC3	P300	6.09	0.44
HeLa-S3	SMC3	POL2	6.00	0.46
HeLa-S3	SMC3	TCF7L2	5.68	0.48
HeLa-S3	SMC3	CFOS	4.97	0.37
HeLa-S3	SMC3	TBP	4.82	0.51
HeLa-S3	SMC3	TAF1	4.71	0.46
HeLa-S3	SMC3	COREST	3.68	0.62
HeLa-S3	SMC3	REST	3.68	0.62
HeLa-S3	SMC3	CHD2	3.58	0.62
HeLa-S3	SMC3	STAT3	3.37	0.53
HeLa-S3	TBP	BDP1	10.37	13.79
HeLa-S3	TBP	RPC155	10.18	11.92
HeLa-S3	TBP	BRF1	3.77	8.57
HeLa-S3	ZNF143	HCFC1	5.02	7.40
HeLa-S3	ZNF143	RAD21	4.32	0.21
HeLa-S3	ZNF143	BAF170	3.54	5.35
HeLa-S3	ZNF143	SMC3	3.23	0.31
HeLa-S3	ZNF143	CTCF	3.21	0.27

Table A4 Presence of ASB events was associated with cobound TFs.

For each ASB dataset, we tested the association between ASB events and binding peaks of its cobound TFs (Fisher test, FDR < 0.05). The p-value and odds ratio of the significant pairs are listed.

Category	Feature Name	Description	Used in model(s)
Motif of the ChIP'ed TF	motif_favor, motif_unfavor	Motif score of two alleles	Full, Seq+DHS, Seq
	Motif_pvalue_ratio	The log ratio of two alleles' binding potential (the p-value of the PWM score against random genome background)	
	Peak_motif_favor, Peak_motif_unfavor, Peak_TFBS_num	Best motif scores within the peak regions for two alleles. The number of predicted TFBS in the peak region	
Position information	Peak_dis	SNV distance to the peak max position	Full, Seq+DHS, Seq
	PWM_position	SNV position in the motif	
Enriched motifs	comotif1_pavlue_ratio, comotif2_pavlue_ratio, …	For each of the five enriched motifs, the log ratio of two alleles' binding potential (the p-value of the enriched motif score against random genome background). We ranked the enriched motifs based on their enrichment within the peak regions.	Full, Seq+DHS, Seq
DHS and 11		The read count at each allele for	

Category	Feature Name	Description	Used in model(s)
histone modifications	DHS_favor, DHS_unfavor, H3K27ac_favor, …	each feature.	Full, Seq+DHS (DHS only)
Cobound TFs	Cobound1, Cobound2, …	Whether the SNV overlap with other TF binding peak regions. We ranked the cobound TFs based on the overlap ratio with heterozygous site binding events.	Full

Table A5. Input features used in three classification models (Seq, Seq+DHS, and Full).

The features are summarized into five categories based on the source or the nature of the data. "Feature names" refers to the features used in Figure 2.5(B) and Appendix Figure A7. Features are explained in the 'Description' column. The last column indicates the models which included corresponding features for training.

A.5 Supplementary data

The compiled heterozygous site binding data can be found in the published paper [52] of Chapter 2. The WGS data of HeLa-S3 cell line used in this research were derived from a HeLa cell line (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000640.v2.p1). Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. This study was reviewed by the NIH HeLa Genome Data Access Working Group.

Appendix B Supplementary material for Chapter 3

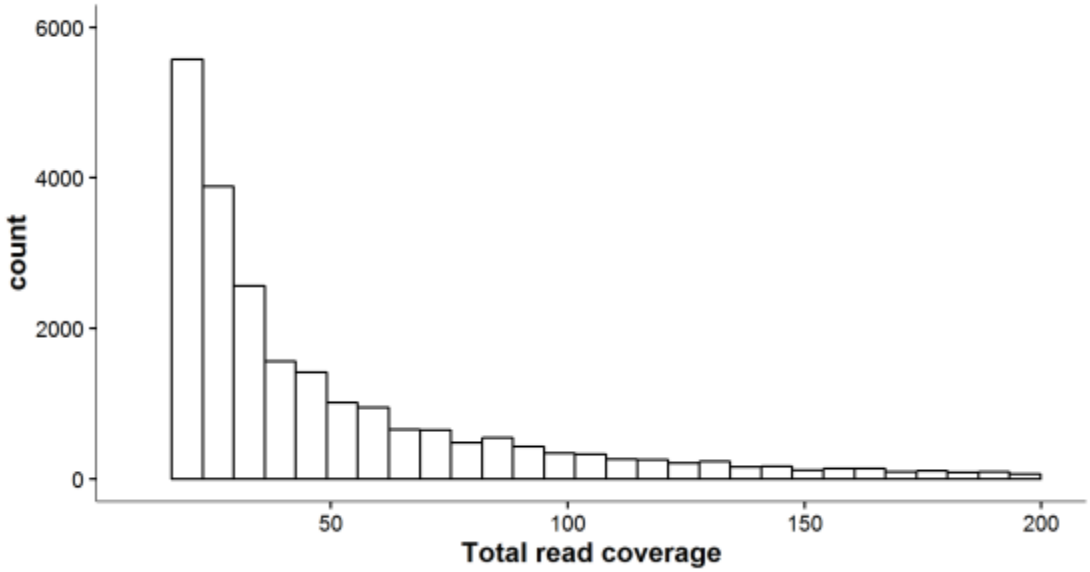


Figure B1. Total read coverage of the heterozygous sites across all the heterozygous site binding events. The sites within read coverage from 10 to 20 represent 41.4% of the sites.

Appendix C Supplementary material for Chapter 4

C.1 Downloading data for LCLs

RNA-seq data for 462 LCLs (individuals) were initially downloaded from the GEUVADIS project [121]. For 445 of them, genotype information was obtained from the 1000 Genomes Project [1]. The individuals covered five populations, including 89 North-Europeans from Utah (CEU), 92 Finns (FIN), 86 British (GBR), 91 Toscani (TSI) and 87 Yoruba (YRI). Because African subjects (YRI) differ substantially from the four European sets, they were excluded from the analysis. The reasons for exclusion included: 1) African individuals exhibit significantly more sequence variations than Europeans [1]; 2) they also exhibit more population-specific differentially expressed genes [121]; and 3) the reference DHS and TF-binding events used in this study are derived from GM12878 cells (an LCL from a European individual).

C.2 Supplementary figures for Chapter 4

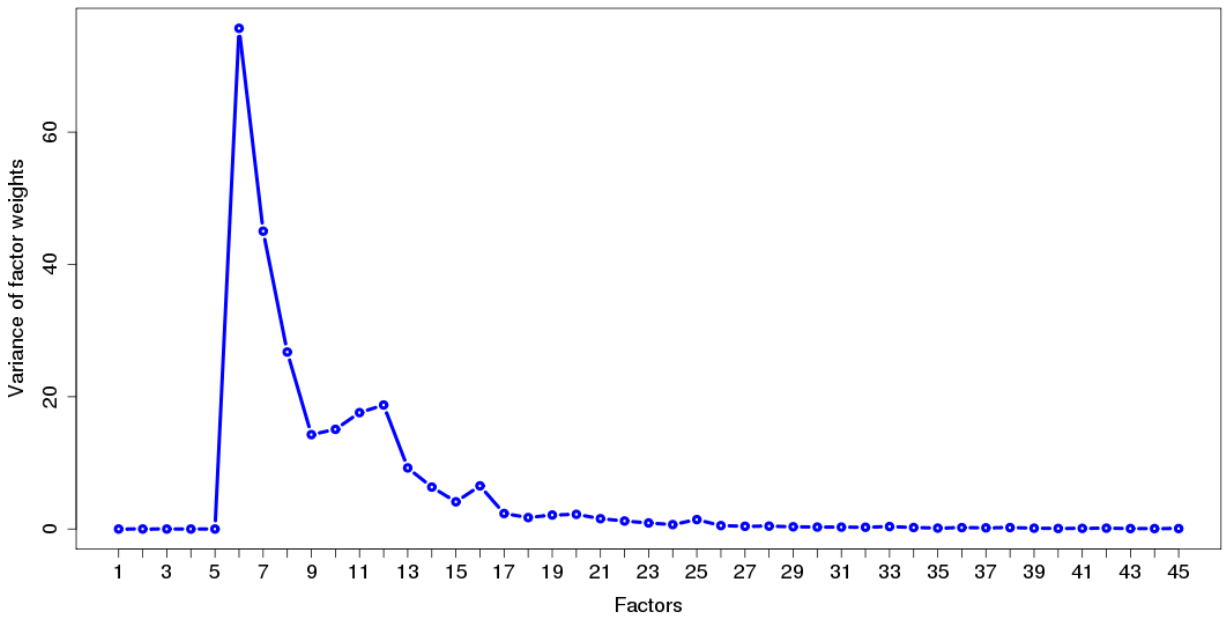


Figure C1. Determine the number of hidden factors in the expression data. We used PEER package to detect the impact of known covariates (four population and one gender factors) and forty potential hidden factors. The natural choice for the number of hidden factors is usually observed as the converged point in the factor variance plot [181]. We chose to remove first 27 factors (The first five known covariates and additional 22 hidden factors) in our expression data.

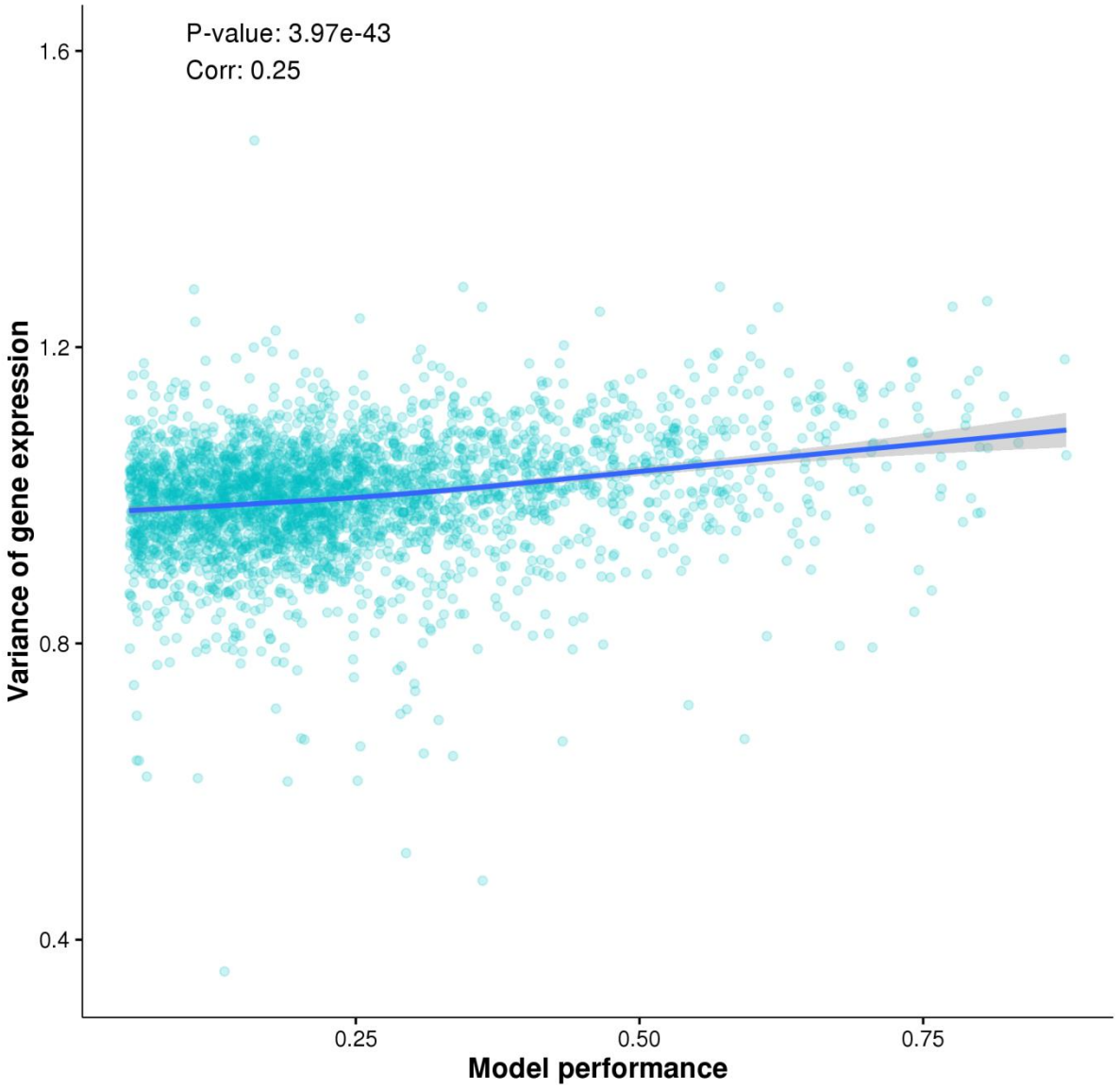


Figure C2. The performances of TF2Exp models are correlated with the variance of gene expressions. Each dot represents one predictable gene. The dot coordinates indicate TF2Exp model performance (x axis) and the variance of gene expression (y axis). We test the correlation between the two axes, and the spearman correlation coefficient and p-value are given on the plot. The blue line shows the general trends drawn by the locally weighted scatterplot smoothing method across all the dots.

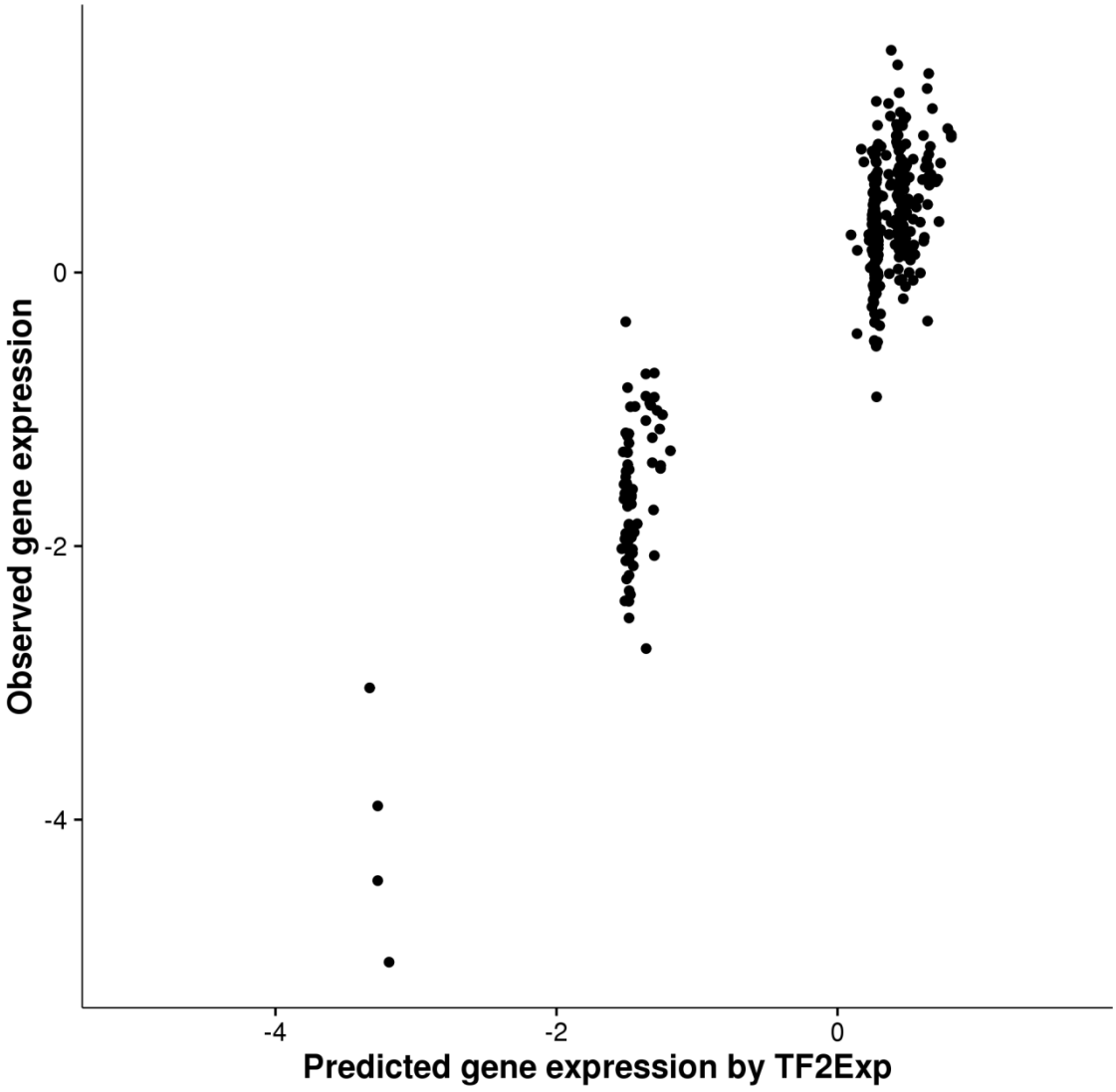


Figure C3. Performance of TF2Exp for BTN3A2 gene in cross validation

Each point represents one CHB tested individual and its coordinates indicate the predicted expression given by TF2Exp model in cross validation (x axis) and the observed expression (y axis).

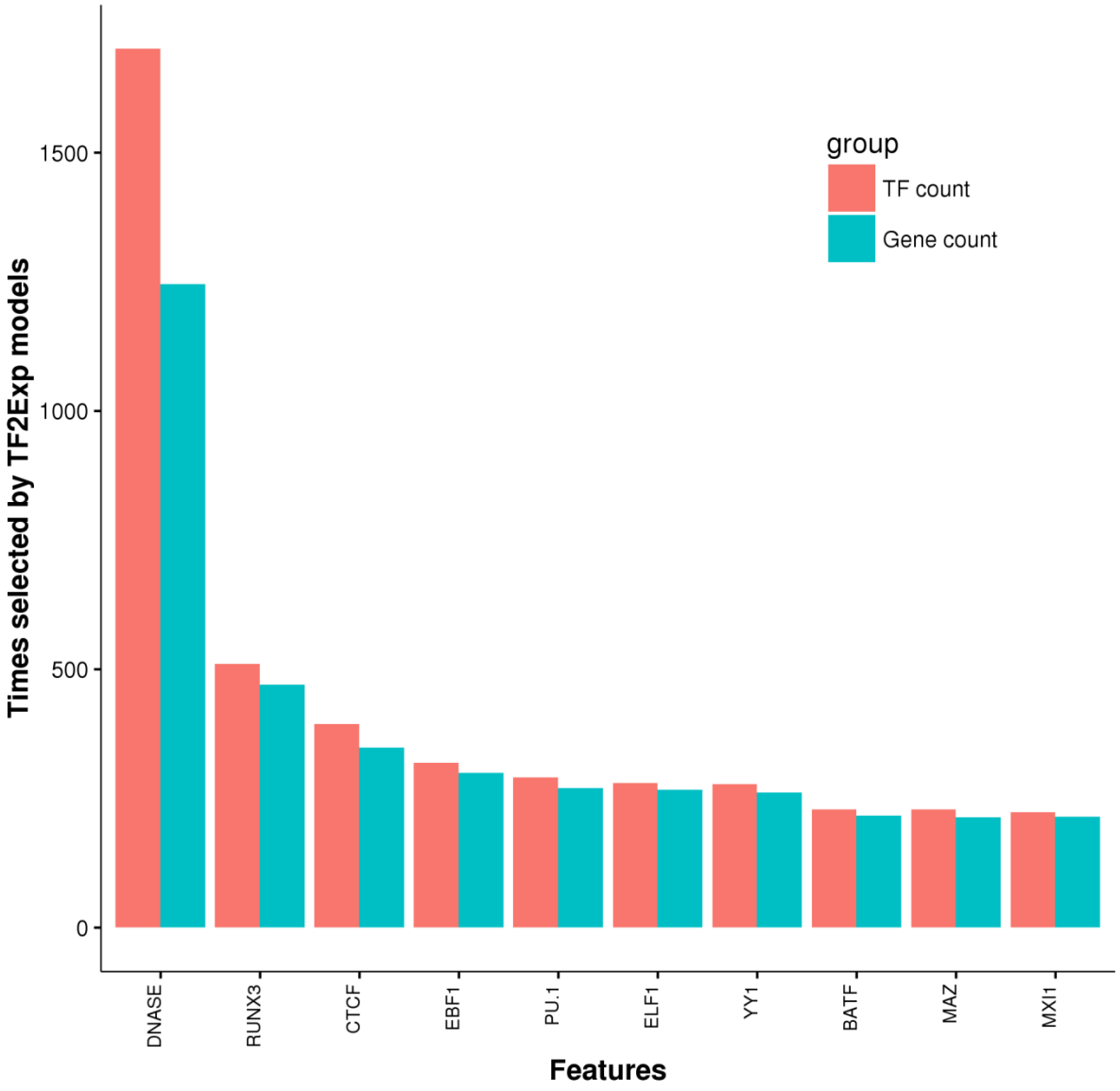


Figure C4 Top 10 TFs whose binding events are the most frequently selected features in TF2Exp models across all the predictable genes.

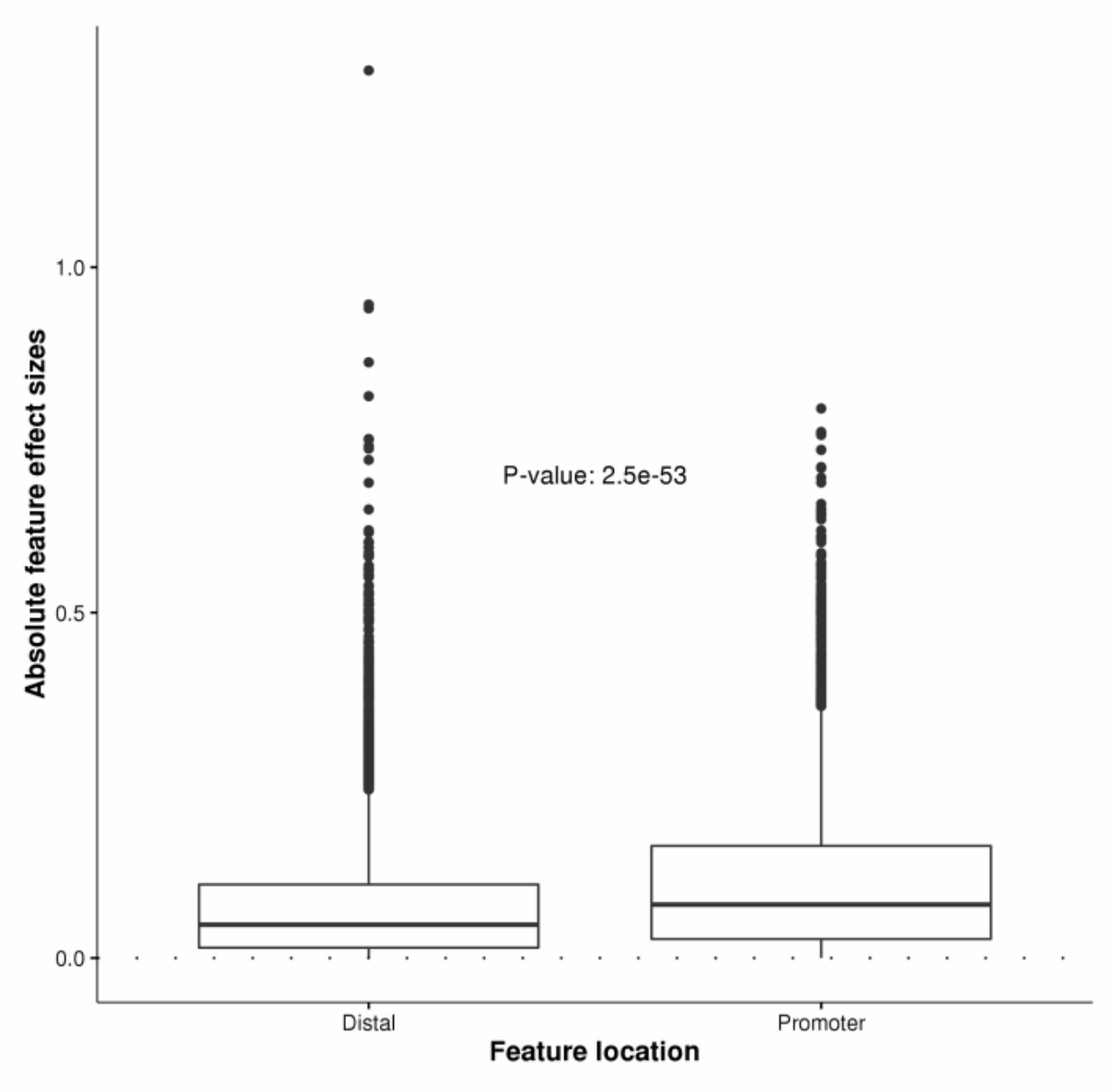


Figure C5. Compare the absolute feature effect sizes of selected TF-binding events at promoter and distal regulatory regions across the all the predictable genes. The labeled p-value indicates the significance for the difference of two groups (Wilcoxon rank-sum test).