Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

TIANLEI XU Date

Statistical and Machine Learning Methods in the Studies of Epigenetics Regulation

By

TIANLEI XU
Doctor of Philosophy

Computer Science and Informatics

Zhaohui Steve Qin
Advisor

Hao Wu
Advisor

Lee Cooper
Committee Member

Peng Jin
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical and Machine Learning Methods in the Studies of Epigenetics Regulation

By

Tianlei Xu
M.Eng. Tongji University, 2012

Advisor: Zhaohui Steve Qin, Ph.D.
Advisor: Hao Wu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2018

Abstract

Statistical and Machine Learning Methods in the Studies of Epigenetics Regulation
By Tianlei Xu

Rapid development of next generation sequencing technologies produces a plethora of large-scale epigenome profiling data. Given the quantity of available epigenome datasets, obtaining a clear and comprehensive picture of the underlying regulatory network remains a challenge. The multitude of cell type heterogeneity and temporal changes in the epigenome make it impossible to assay all epigenome events for each type of cell. Computational model shows its advantages in capturing intrinsic correlations among epigenetic features and adaptively predicting epigenome marks in a dynamic scenario. Current progress in machine learning provides opportunities to uncover higher level patterns of epigenome interactions and integrating regulatory signals from different resources. My works aim to utilize public data resources to characterize, predict and understand the epigenome-wide regulatory relationship. The first part of my work is a novel computational model to predict *in vivo* transcription factor (TF) binding using base-pair resolution methylation data. The model combines cell-type specific methylation patterns and static genomic features, and accurately predicts binding sites of a variety of TFs among diverse cell types. The second part of my work is a computational framework to integrate sequence, gene expression and epigenome data for genome wide TF binding prediction. This extended supervised framework integrates motif features, context-specific gene expression and chromatin accessibility profiles across multiple cell types and scale up the TF prediction task beyond the limits of candidate sites with limited known motifs. The third part of my work is a novel computational strategy for functional annotation of non-coding genomic regions. It takes advantage of the newly emerged, genome-wide and tissue-specific expression quantitative trait loci (eQTL) information to help annotate a set of genomic intervals in terms of transcription regulation. This method builds a bridge connecting genomic intervals with biological pathways and pre-defined biological-meaningful gene sets. Tissue specificity analysis provides additional evidence of the distinct roles of different tissues in the disease mechanisms.

Statistical and Machine Learning Methods in the Studies of Epigenetics Regulation

By

Tianlei Xu
M.Eng. Tongji University, 2012

Advisor: Zhaohui Steve Qin, Ph.D.
Advisor: Hao Wu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2018

Table of Contents

List of Tables

List of Figures

Chapter 1
Epigenomic feature prediction from high-throughput data

1.1 Introduction

1.1.1 Epigenomic research and high-throughput data

Epigenomics is the study on the regulation of gene expression. Literally it refers to the specific mechanisms that control the outcome of expression result without altering the DNA sequence, which is the static format of encoded genetic information that doesn't change dynamically through development of different cells. Epigenetic factors regulates the chromatin structure and gene expression in eukaryotes. Study on mechanisms of epigenetic regulation in diverse cell contexts is the key to a better understanding of the basis of biological processes and diseases. The epigenetic regulation are carried out by multiple factors, including DNA methylation(Lister et al., 2009), histone modifications (Barski et al., 2007; Heintzman et al., 2009) and the chromatin accessibility(Thurman et al., 2012). The activities of these epigenetic marks are highly tissue and cell type-specific. Assessing the distribution of epigenetic marks on whole genome scale is proved to be a powerful technique to infer functional roles of each member.

High-throughput technologies have been developed in fast pace in the recent decades. A plethora of sequencing platforms (Bentley et al., 2008; Harris et al., 2008; Margulies et al., 2005; McKernan et al., 2009; Shendure, 2005) brings the cost of reading the characters of genetic samples to a unprecedented low level. Ideally, a full range of epigenetic profiles

should be investigated for each cell context. Multiple research consortiums, including NIH Roadmap Epigenomic Consortium(Bernstein et al., 2010; Roadmap Epigenomics Consortium et al., 2015), ENCODE (Bernstein et al., 2012), have made considerable endeavors towards such goal, and a rich collection of epigenetic profiles became available thanks to these efforts. TCGA is trying to do that. However it's prohibitively expensive and cannot be practically applied in large-scale study and clinical settings. Thus, computational prediction is helpful to provide an alternative when the epigenetic data are unavailable. With the accumulation of high-quality epigenomic profiling data, and the development of modeling methodology, in silico study on epigenomics has been gaining popularity ever since. Calls new statistical and computational methods to be developed to address the many challenges analyzing epigenomics data (Qin et al., 2016)

1.1.2 Epigenomic features in gene regulation

In general, the genetic features can be categorized into two classes: the static features, and the dynamic features. The static features are based on DNA sequence, thus they do not change across cell types within one individual. These features only need to be measured once, then it can be applied among different cells. Dynamic features are cell type-specific features. They are the direct reason for a diverse range of function and shapes in the cell population from the same individual. The heterogeneity of cells result in the complexity of characterizing all the dynamic features, given that we currently ignore the impact of genetic variants. A list of members from both classes are shown below.

Static Features:

- DNA Sequence composition (GC content, k-mer, etc.)
- DNA motifs (TF specific)
- Gene annotation (including gene itself and regulatory regions)
- CpG island
- Sequence conservation
- Shape of DNA

Dynamic Features:

- Gene expression
- DNA methylation
- Histone modification
- Protein-DNA interaction
- Chromatin structure (spatial organization, open/close chromatin)

In silico prediction is possible due to the intrinsic correlations among features. Although the observed correlation among features doesn't imply any causal relationships, it is still important to mine knowledge from vast data for the most predictive or relevant features, which would bring insights to further facilitate research on mechanism of transcription regulation.

1.1.3 Prediction methods

Predictive models have been proposed all the time along with the progress of research in the study of epigenetics. At least two benefits from models are noteworthy in this co-evolutionary process. The first is the modeling of underlying mechanism of epigenetics regulation. Without any prior knowledge on the function of epigenetic marks, generating

rules with statistical methods from observation is the first step to understanding the true underlying mechanism. Predictive models is the automation of representation for generalized rules we learned from data, and it can help to validate the discovery by testing models on different scenarios. The second is the benefits to support decision making. Before the true biological mechanism is validated, computational models provide useful tools to generate complex rules based on existing information that could bring profound improvement on the research efficiency. In this work, both supervised models and unsupervised models will be discussed. While it is reasonable to make a safe assumption that for predictive models, supervised learning should be the dominant type of method, I will show in the following sections that unsupervised models also plays an important role in biological studies. The heterogeneity of sample context makes it necessary to discover the inner structure of the observed data before dumping all the available data into a simple and unified machine learning framework. In addition, due to the raw data might not be applicable to generate well defined features, many models are hybrid computational frameworks that perform rigorous feature engineering before proceeding to build predictive models. In the following sections, I will provide a comprehensive survey of existing predictive models in the study of epigenetics. The methods are grouped based on the target of the prediction. Within each group, the detailed nature and characters of target marker, together with the rationale behind the selection of features and models will be discussed in detail.

1.2 Prediction of protein binding

1.2.1 Rationale for transcription factor binding prediction using epigenetic profiles

The task of predicting binding sites for regulatory proteins, or transcription factors (TFs), is necessary, due to the fact that experimental methods can only determine the true binding sites of one type of TF, under one condition (tissue, cell, treatment/disease, etc.) at a time. It is impossible to enumerate all the combinations between every TF and cell condition to sequence. Thus, computational methods are becoming popular in this area, using existing data and knowledge to generate rules of TF binding, and trying to impute the binding profile in a new cell condition that no data are available.

In general, several factors participate in the regulation of TF binding. These factors include: DNA sequence motifs, chromatin accessibility, Histone modification and methylation status. Previous to the comprehensive study of dynamics in TF binding regulation, DNA sequence motif is the most widely used feature to predict TFBS in silico. TF binding motif can be found in database such as JASPAR (Sandelin, 2004) and factorbook (J. Wang et al., 2013). Motif based methods suffer in performance, due to the fact that static genome sequence cannot be the only determinant factor for the cell-type-specific dynamic events, such as TF binding. It should be noted that the dynamics of TF binding is not simply a repository of conserved binding sites being switched on and off by epigenetic landscape. For one TF, the sequence patterns in different cells may vary (Arvey, Agius, Noble, & Leslie, 2012). A recent study also shows that on individual level, the repository of TF binding activities may be affected by one's genetic variation (Barrera et al., 2016). We will not discuss existing works on sequenced based methods. Instead we will discuss how sequence-based information is combined with other features in prediction.

1.2.2 Prediction methods using histone modifications

Histone marks were found in close correlation with regulatory activities in the human genome (Heintzman et al., 2007). Different histone marks were reported associated with different regulatory elements, such as open chromatin, promoter, enhancers etc. (Heintzman et al., 2007; Schones et al., 2008).

Whitington et al. (Whitington, Perkins, & Bailey, 2009) firstly used histone mark H3K4me3 as filters for TFBS predictions based on motif only. Combined with other features, such as distance to known TSS and sequence conservation, their work reduced the false positive rates in TFBS predicting result. He H.H et al. used differential H3K4me2 ChIP-seq signals to measure nucleosome positioning, followed by motif analysis to predict TF binding dynamics(H. H. He et al., 2010). Talebzadeh used the composition of different histone marks within neighboring nucleosomes as predicting features. Then it is combined with PWM to fit in a logistic regression classifier to predict TF binding (Talebzadeh & Zare-Mirakabad, 2014). Stephen A. Ramsey et al used local-minima of Histone Acetylation ChIP-seq signals (valley score) combined with motif scanning score to predict TF binding sites . Binding score is assigned by a weighted sum of score from different features. Supervised training is performed to infer model parameter and hold-out test to evaluate model (Ramsey et al., 2010). Won et al used a more comprehensive set of histone features to train HMM models in a supervised fashion to predict TFBS (Won, Ren, & Wang, 2010). This method, Chromia, takes both PSSM signal and binned histone modification signal as input, and fit a three-state mixture Gaussian model. It could also implement conservation score and genomic features that treated promoter sequence and enhancer differently. But the features used inside less well characterized enhancer regions are harder to be generalized in prediction, as shown in the worse performance in enhancer TFBS

predicting result. Another interesting finding is the heterogeneity of histone contribution. There is no histone mark that are constantly associated with TFBS. H3K4me3 are the most frequent strong predictor. The best performed TF, E2f1, shows the strongest association with H3K4me3. Interestingly, CTCF behaves differently from other TFs listed as major results in this paper. CTCF prediction have bad performance using Chromia, which is almost similar to baseline method. Adding conservation score might sabotage promoter TFBS prediction, however will increase enhancer TFBS prediction for CTCF. These results suggest that histone modification is definitely not the global indicator of TF binding signals. Histone marks that are associated with chromatin states might help improving relevant TFBS.

It is interesting to notice that among a variety of histone marks, only a few are used, selected, or proved to be relevant with TF binding regulation. This might suggest their association with other confounding factors, rather than a direct binding signal. These factors can be open chromatin (HAc, (Ramsey et al., 2010)), nucleosome distribution (H3K4me2, (H. H. He et al., 2010)), transcription (H3K36me3, (Won et al., 2010)), promoter (H3K4me3) or enhancers (H3K4me1). Based on this idea, Ji et al. (Ji, Li, Wang, & Ning, 2013) use histone marks to define genomic categories and made TFBS prediction based on this information. The histone marks serve as a feature space in which the complete epigenome environment could be projected.

1.2.3 Prediction methods using chromatin accessibility

Ever since the chromatin accessibility data became available via large epigenomic projects, this topic is dominated by predicting methods using chromatin accessibility information as

features. It is intuitively ideal, that the direct measurement of open chromatin structure is the straight-forward evidence for a protein-DNA interacting events. There are in general two big categories of methods using chromatin data: bin-based methods and candidate site based methods. Bin-based methods include DNase2TF (Sung, Guertin, Baek, & Hager, 2014) and HINT (Gusmao, Dieterich, Zenke, & Costa, 2014). This class starts with searching for footprint signature of TFBS shown on DNase hypersensitive site sequencing, where a short range of DNase data is depleted from two peaks of cleavage sites. This footprint is usually enclosed within signatures of histone modifications. HMM model was adopted to recognize this type of local-dependency relationship in HINT and Chromia. The other class are based on candidate sites predicted using known TF motifs. Then the chromatin profile centered at these candidate sites are analyzed to detect potential TF binding sites. This class contains both supervised and unsupervised learning methods. In unsupervised methods, CENTIPEDE (Pique-regi et al., 2011), FootprintMixture (Yardimci, Frank, Crawford, & Ohler, 2014) and PIQ (Sherwood et al., 2014)are three representative works. These methods, including Romulus (Jankowski, Tiuryn, & Prabhakar, 2016) and MOCAP (Chen, Yu, Carriero, Silva, & Bonneau, 2017), capturing differences between binding sites and non-binding sites, then phase out the labels using prior knowledge. This type of methods are useful when there is no training data available for supervised learning methods, and models can learn the class structure directly from data. Other methods use a variety of representations of DNase profile, some combine with additional epigenetic marks as prior (Cuellar-Partida et al., 2012)for this task, and train different types of models, including SVM (Arvey et al., 2012; Quach & Furey, 2017) or random forest (Kuang, Ji, Boeke, & Ji, 2017; S. Liu et al., 2017).

One problem associated with chromatin-based predicting methods are the sequence bias issue of the DNase foot printing experiments, which might lead to false detection of enriched motif resulted from technique noise, rather than sequence signal of TF binding. Gusmao et al. discussed this issue in detail in a recent review work (Gusmao, Allhoff, Zenke, & Costa, 2016), in which details of this issue is analyzed with results from multiple methods.

1.2.4 Other methods for TFBS prediction

The prediction of TFBS has been an active field for nearly a decodes, and new predictive features has been adding to this category continuously. Methods using methylation profile (Xu et al., 2015) or DNA shape features(Ma, Yang, Rohs, & Noble, 2017) have been proposed in addition to popular epigenetic marks. In addition, recent popularity of deep learning technology brought applications in TFBS prediction as well. DeepBind (Alipanahi, Delong, Weirauch, & Frey, 2015) used convolutional neural network (CNN) to detect motif-like DNA sequence kernels, and using them as input features in a feed-forward neural network to predict binding affinity of proteins. FactorNet (Quang & Xie, 2017) added an additional layer of recurrent neural network (RNN) to model the spatial dependency of feature signals. A full list of TFBS prediction methods reviewed are shown below.

Table 1.1 Predition methods for TFBS

Publication	Features	Method	Software

Whitington et al. (2009)	Histone ChIP-seq	PWM scan followed by filtering Chromatin features	N/A
He et al. (2010)	Nucleosome-resolution histone ChIP-Seq	Dynamics of nucleosome occupancy and motif	N/A
Won et al. (2010)	Histone ChIP-seq	HMM	Chromia
Ramsey et al. (2010)	Histone acetylation ChIP-seq, nucleosome occupancy, genome	Weighted sum of scores.	RamseyHAc2010
Pique-Regi et al. (2011)	DNase I + genome	Two-component mixture model, EM. unsupervised	CENTIPEDE
Cuellar-Partida et al. (2012)	DNase I, histone ChIP-seq	Epigenetic data as prior, use motif to predict.	FIMO, part of MEME
Arvey et al. (2012)	Histone ChIP-seq + DNase I	SVM	N/A
Ji et al. (2013)	Nine histone ChIP-seq	PCA-type unsupervised learning.	dPCA
Gusmao et al. (2014)	DNase + histone	HMM	HINT
Sung et al. (2014)	DNase + Motif (4-mer)	Tests based on counts	Dnase2TF
Yardimci et al. (2014)	DNase + bias adjustment	2-component mixture model.	FootprintMixture
Sherwood et al. (2014)	DNase (magnitude + shape) around motif match sites	2-component mixture model; Gaussian process model for DNase reads	PIQ

Kahara et al. (2014)	DNase + motif score	Logistic regression + greedy backward feature selection	BinDNase
Xu et al. (2015)	Methylation + genomic features	random forest + 2-component mixture model for methylation	Methylphet
Alipanahi et al. (2015)	DNA sequence (using binding array data as target)	CNN	DeepBind
Quach and Furey (2016)	DNase (profile: mean and slope, centered at motif)	SVM	DeFCoM
Jankowski et al. (2016)	DNase (shape, on motif matched sites)	2-component mixture model	Romulus
Liu et al. (2017)	DNase (footprint score defined by counts) + genomic features	Random Forest	BPAC
Ma et al. (2017)	DNA sequence + shape kernel	Support vector regression	Sequence-shape
Kuang et al. (2017)	Histone + DNase, on known motif matched sites	Random Forest	DynaMO
Chen et al. (2017)	ATAC-seq	3-component mixture model, EM; Negative-binomial; unsupervised	Mocap
Quang et al. (unpublished)	DNase	CNN+RNN	FactorNet

1.3 Prediction of enhancer

1.3.1 Diversity of definition of Enhancer

Enhancer prediction is difficult, potentially decided by the lack of a unified standard to define what is the gold standard for an enhancer-gene pair. The question about what is an enhancer is unquestionably more important before we ask how to predict them. By far, predicting models rely on experimentally validated enhancers from public data resources, such as FANTOM (Andersson et al., 2014) or ENCODE (Bernstein et al., 2012), where the spatial chromatin organization that brings the regulatory enhancer to a distal promoter region of a gene is considered as the validated evidence of an enhancer. However, this approach is not applicable to all the cells/tissues due to the cost. Alternative standards are proposed to use epigenetic features that are faithfully present at enhancer region as the indirect gold standard. Thus it will be contra-intuitive that some epigenetic marks are used as the definition of enhancer, rather than the features to predict them, or at least, the border line between features and gold-standards is tricky, and need to be carefully specified when a study of enhancer prediction is conducted.

With limited annotation resource, the task of predicting enhancers are necessary yet non-trivial. Unlike well-annotated gene/TSS/promoter, this task only became possible with the accumulation of TFBS/DHS/Histone data available. Histone modifications are considered associated with enhancer, acting as the markers to guide different transcription factors to form regulatory chromatin loops. These modifications include H3.3 and H2AZ (Jin et al., 2009), H3K4me1 and H4K27ac (Cotney et al., 2012; Heintzman et al., 2007; Koch et al., 2007). The latter is specifically studied as the marker for poised enhancer (Creyghton et al., 2010; Rada-Iglesias et al., 2011). Nucleosome positioning is also a key factor to define enhancers (H. H. He et al., 2010), which could potentially be the result of chromatin opening. Also, transcription factors are essential components of enhancer activity as well.

Multiple studies showed the key role of TFs as the marker of enhancers, such as p300/CBP (Blow et al., 2010; Ghisletti et al., 2010; May et al., 2012; Visel et al., 2009) or a combination of multiple TFs (Cheng et al., 2012; A. He, Kong, Ma, & Pu, 2011; Yip et al., 2012; Zinzen, Girardot, Gagneur, Braun, & Furlong, 2009). In prediction methods, common practice is to overlap predicted "enhancer" with P300 + DHS + several TFBS known to be associated with enhancer. Overlapping with TSS is regarded as negative gold-standard, especially for methods using histones as predictors, since promoters share a large portion of histone features with enhancers.

1.3.2 Challenges of enhancer prediction

As described from previous section, the definition of enhancer positive and negative class is up to the researchers. Especially the negative set of enhancers – it is noteworthy that one negative sites in one tissue might be an active enhancer in another tissue, the context matters at all time. It is therefore important to choose negative set in the training set carefully, after all, the performance of the model can only be as good as the training data quality that are fed to the model. For example, transcription start sites share a considerable amount of common features with enhancers, thus the negative sites must include TSS regions should the user differentiate them from distal regulatory regions.

Tissue specificity is another issue to be considered. Just like gene expression and other epigenetic profiles, enhancer is highly cell-type specific (Paige et al., 2012; Wamstad et al., 2012). The enhancers regulating developmental genes should not function in developed tissues in principle. In accordance, the relevant features, such as developmental TFs like SOX2/OCT4/NANOG in embryonic cells, will function differently with enhancer

regulation in other tissues. Thus, the dilemma of prediction methods emerges: if the model cannot generate rules of enhancer regulation across different cell types, then the model is in general useless. Within cell prediction is not indeed necessary, given the fact that whole genome profiling techniques will take a snapshot of the full range of genome.

Another challenge of enhancer prediction is the diversity of enhancer itself. Studies have shown that the enhancers are highly heterogeneous(Bonn et al., 2012; Zentner, Tesar, & Scacheri, 2011). Thus, for the prediction task of enhancer, it is essential for the selection of training data to collect a comprehensive set of representing training samples, both for positive classes and negative classes.

1.3.3 Tools for enhancer prediction

Before epigenetic marks were adopted for predicting enhancer, DNA sequence was investigated as the potential predictor for enhancers (Heintzman et al., 2007), and this information has been used as enhancer predictor for a long time. K-mer based methods are always coupled with SVM classifier, and new representation of k-mers allowing gaps was introduced as well (Ghandi, Lee, Mohammad-Noori, & Beer, 2014; Lee, Karchin, & Beer, 2011). Other variants based on motifs (Taher, Narlikar, & Ovcharenko, 2012), DNA local structure (B. Liu, Fang, Long, Lan, & Chou, 2015) were proposed as well. As more epigenomic data become available, tools like (Jia & He, 2016) that based only on sequence are not as popular as before.

Among all the epigenomic marks, histone modifications are the most relevant to enhancer. Firpi et.al developed a tool for enhancer prediction in early stage of enhancer prediction (Firpi, Ucar, & Tan, 2010). They used neural network to model the histone modifications

as features at enhancer regions. It was considered as the standard work that many following works adopt the same setting of task, including the size of negative sets, validation methods and so on. Non-linear effects among different histone modifications were modeled by SVM (Miranda-saavedra, 2012), random forest (Rajagopal et al., 2013), adaboost (Lu, Qu, Shan, & Zhang, 2015) or combined with ensemble learning framework (Kleftogiannis, Kalnis, & Bajic, 2014). The current trend to predict enhancer is to utilize all the available epigenetic profiles and capture the inter-relationships by complex models. Such methods combine histones with chromatin accessibility (Erwin et al., 2014; F. Liu, Li, Ren, Bo, & Shu, 2016) or methylation (Y. He et al., 2017). A comprehensive list to these tools are shown below, the length of which will continue to grow due to the increasing amount of epigenome data becoming available.

Table 1.2 Prediction methods for enhancer

Publication	**Enhancer definition/control**	**Features Used**	**Method**	**Software**
Heintzman et al. (2007), Nature Genetics	P300/random	DNA sequence	correlation	
Firpi et al. (2010), Bioinformatics	74 validated from Heitzman et al. 2007	histones	ANN (TDNN)	CSIANN
Lee et al. (2011), Genome Research	P300/random	DNA sequence (k-mer with k=3~10)	SVM	Kmer-svm
Taher et al. (2012),	Validated/random	TF motifs	LASSO regression	CLARE

Bioinformatics				
Miranda-saavedra et al. (2012), NAR	P300 distal to TSS/random	histone	SVM + genetic algorithm	ChromaGenSVM
Rajagopal et al. (2013), Plos Comp. Bio	p300 overlapping with DHS, distal to TSS/validated	24 histones	Random forest	RFECS
Ghandi et al. (2014), Plos Comp. Bio	p300 in mouse embryonic/ random	Gapped-kmer	SVM	Gkm-svm
Erwin et al. (2014) Plos Comp. Bio.	VISTA enhancer/tissue-specific non-enhaner validated	step 1: histone, TFBS, Dnase/FAIRE, conservation, motif; step 2: histone, p300	linear SVM as step 1; multiple kernel learning in step 2	EnhancerFinder
Kleftogiannis et al. (2015), NAR	ENCODE validated enhancer / random	Histone, sequence	Ensemble, with SVM as base classifier and ANN as final output classifier	DEEP
Liu et al. (2015), Bioinformatics	Validated enhanver/non-enhancer	Sequence(k-mer), DNA local structure	SVM	iEnhancer-2L
Lu et al. (2015), Plos ONE	P300, DHS; distal to TSS/random	Histone (shape of profile in addition to intensity)	adaboost	DELTA
Liu et al. (2016)	H3K27ac peaks; multiple	9 category: histone modifications, 27	deep learning(DNN) +HMM;	PEDLA

Scientific Report	filters (distal to TSS, etc);	TFs and cofactors, 15 chromatin accessibility, transcription, RRBS, CpG islands, evolutionary conservation, sequence k-mers, motifs (TFBS)	Iteratively train through cell types	
Jia et al. (2016) Scientific Report	Strong or weak enhancer / non-enhancer	400bp: bi-profile bayes(BPB) (similar to 200bp positive PWM and 200bp negative PWM; Nucleotide frequency; pseudo-nucleotide frequency:3-mer frequency;	SVM	EnhancerPred
He et al. (2017) PNAS	P300/random + promoters	Histones, methylation	Random forest	REPTILE

1.4 Prediction of DNA methylation

DNA methylation is an important component among all epigenetic marks. A variety of biological processes including development and differentiation, have methylation changes as a signature. The majority of DNA methylation is on the Cytosine of CG dinucleotides(5mc), or CpG sites, although other types of DNA methylation (non-CG methylation) exists as well. There are also variation of methylation types, such as hydroxymethylation (5hmC).

In early studies when high throughput sequencing technologies are not yet available, methylation level of CpG sites are investigated together with DNA sequence pattern using methylation specific restriction enzymes (Rollins et al., 2006), and prediction models aim to infer methylation levels based on the limited rules generated from sequence. MethDB (Grunau, 2001) was the popular choice of the training resource then. Multiple works (Bhasin, Zhang, Reinherz, & Reche, 2005; Das et al., 2006; Fang, Fan, Zhang, & Zhang, 2006) were proposed to use sequence-derived features for methylation prediction on limited CpG sites. However, an intrinsic problem with these initial exploring works is that epigenetic marks, such as DNA methylation is cell/tissue type specific, thus, using only static genomic DNA sequence is not sufficient to capture its dynamic profile, or at least, this strategy can only be applied within specific genomic regions containing functional elements (Whitaker, Chen, & Wang, n.d.).

High-throughput sequencing technologies enable researchers to profile methylation landscape in larger scale with lower costs. Microarrays such as the Illumina Infinium Methylation450k and MethylationEPIC array, are designed to cover CpG sites in important regulatory genomic regions. Reduced representation bisulfite sequencing (RRBS) extended the range of research to the genomic scale that are enriched with CpG sites(Meissner et al., 2005). Whole-genome bisulfite sequencing (WGBS) covers CpG sites across the whole genome. Prediction of DNA methylation has some distinct features compared to predictions on other epigenetic marks. The majority of tools for this task are to perform imputation across the whole genome (Angermueller, Lee, Reik, & Stegle, 2017; Fan, Huang, Ai, Wang, & Wang, 2016; Y. Wang et al., 2016; Zeng & Gifford, 2017; Zhang, Spector, Deloukas, Bell, & Engelhardt, 2015), usually within the same cell/tissue, rather

than transferring known mechanism to a new cell/tissue context. This is largely the result of DNA methylation measurement techniques: array-based methods are cost-efficient, but only cover a portion of all the CpG sites; on the other hand, WGBS method performs a genome-wide survey of all CpG sites, but the expenses limits its application, also, inaccurate measurement of methylation level on low coverage CpG sites occurs within WGBS data. Thus, computational imputation methods evolves to address the tradeoffs.

With the growing numbers of epigenome profiling data available, imputation can be done without initial probing of limited CpG sites. Predicting methods using relevant markers to predict methylation became possible. Zhang et al. integrate sequence signatures together with other cell-type specific markers, including DNase Hypersensitive Sites(DHS) and transcription binding sites (TFBS) to predict methylation level (Zhang et al., 2015). A different group introduced the other possibility to consume chromatin topological features for the same task (Y. Wang et al., 2016). A recent method brings in histone modification to the already rich collection of features for the task of methylation prediction (Zou et al., 2018). In addition to prediction of methylation level for CpG sites in general, Zeng et al. proposed to predict the effect of sequence variants on CpG methylation (Zeng & Gifford, 2017), which is similar to the application of DeepBind (Alipanahi et al., 2015). Also, imputation methods might find their new application in single-cell data, due to the high missing values in single-cell methylation data. Angermueller et al. proposed a deep learning framework, using convolutional neural net (CNN) to capture sequence features, and recurrent neural net (RNN) for spatial dependent relationships among CpG sites to impute missing data for single cell methylation sequencing results (Angermueller et al., 2017).

A list of the methods in methylation prediction is shown in Table below.

Table 1.3 Prediction methods for methylation

Publication	**Methylation Type**	**Features Used**	**Method**	**Software**
Bhasin, et al. (2005), FEBS	CpG sites from MethDB	DNA sequence	SVM	Methylator
R. Das, et al. (2006), PNAS	CpG sites, differentiate in/out CGI	DNA sequence	SVM	HDFinder
Fang, et al. (2006), Bioinformatics	CGI	GC content, DNA motifs,	SVM	MethCGI
Whitaker et al. (2015), Nature Method	DNA methylation valleys	DNA sequence (motif)	LASSO feature selection + random forest prediction	epigram
Zhang, et al. (2015), Genome Biology	WGBS, 450K	DNA sequence,(composition, recombination rate, evolution rate), 450K data, DHS site, TFBSs	Random Forest	[not available]
Wang, et al. (20156), Genome Biology	RRBS, Hi-C	DNA sequence (composition, k-mer) , both local and remote based on Hi-C data	DNN (stacked denoising autoencoders)	deepmethyl
Fan, et al. (2016), Genomics	WGBS, 450K	DNA sequence (k-mer), 450K data	Random Forest	[not available]
Zeng, et al. (2017), Nucleic Acid Research	RRBS, meQTL	DNA sequence	CNN of Keras	CpGenie

Angermueller, et al. (2017), Genome Biology	Single-cell RRBS	DNA sequence, neighboring CpG	CNN + RNN	DeepCpG
Luli S Zou, et al. (2018), bioRxiv	WGBS, EPIC,	ATAC-seq, Histone, TFBS, Genomic Features(CGI, GC content, recombination rate)	XGBoost	BoostMe

1.5 Prediction of spatial chromatin structure

In this category, the prediction methods vary a lot due to the complexity of characterizing chromatin structure. Different experimental techniques provide diverse views of the spatial organization of chromatin.

A/B compartments are one way to divide the genome based on chromatin structure. Interactions between loci within one compartment are independent to the other. The A compartment was found to be associated with open chromatin while the B compartment with closed chromatin. The simple yet inspiring assumption behind the works to predict chromatin structure is that distal but interacting loci turn to harbor similar epigenomic features. Fortin et al. (Fortin & Hansen, 2015) used this strategy to reconstruct A/B compartment using methylation data and other types of epigenomic marks across multiple cell lines. Genomic loci with high correlation of these marks are predicted to be within the same compartment.

Similarly, Zhu et al. (Zhu et al., 2016) developed a novel strategy to detect spatial chromatin interaction structure measured by capture-C experiments, using 1D epigenomic data. They collected epigenomic marks including chromatin accessibility, histone

modifications and gene expression levels from 5 different tissue types. Then correlation of distal loci are measured based on tensorized epigenomic marks, and significantly associated loci are called by permutation. Huang et al. (Huang, Marco, Pinello, & Yuan, 2015) adopted a different way to study the structural features by defining interaction hubs and train a machine learning classifier based on epigenomic marks. In addition to the efforts to utilize epigenomic marks, molecular thermal simulation has been applied to explore the structural feature of chromatin as well; Brackley et al. (Brackley et al., 2016) proposed to use chromatin accessibility data to infer TF binding sites, which could serve as the interaction points to form protein bridge. Then polymer modeling were applied to simulate the thermal motion of the chromatin fiber to detect potential local interaction structures based on inferred protein bridging sites.

A recent study extend the prediction target from the looping structure of chromatin to accessibility prediction (Jung, Angarica, Andrade-Navarro, Buckley, & Del Sol, 2017). Transcriptome data is used to infer the chromatin landscape within gene-regulatory regions. The intent to predict high-cost signals from low-cost transcriptome data sounds rational, however, the limitation is obvious as well: only with known regulatory relationships between genes and regulating genomic regions, the availability of which across multiple cell/tissue types is questionable, can the prediction be made faithfully.

Table 1.4 Prediction methods for chromatin structure

Publication	**Targeted Chromatin**	**Features Used**	**Method**	**Software**
Fortin et al. (2015) Genome Biology	A/B compartment	450K Methylation, microarray, DHS,	Eigen vector of features + Correlation	(R script provided)

		scATAC-seq, scWGBS		
Huang et al. (2015), Genome Biology	chromatin interaction hubs and topologically associated domain (TAD) boundaries.	Hi-C, histones, DNA sequence	Bayesian additive regression tree	N/A
Zhu et al. (2016), Nature Communication	Spatial association within TADs	Histone, DHS, RNA-seq	Tensor vector correlation + permutation	EpiTensor
Brackley et al. (2016), Genome Biology	Local folding/interaction map, on Alpha/beta globin loci in mouse erythoblasts	DHS + TF motifs	Polymer model	N/A
Jung et al. (2017) Scientific Report	Chromatin accessibility (ENCODE DNase-seq)	Transcriptomic data (ENCODE RNA-seq)	hierarchical random forest	ChromAccPrediction

1.6 Prediction of gene expression.

Gene expression prediction is the downstream target in the whole cascade of regulatory network. All the other elements, including chromatin accessibility, histone modification and transcription factor binding, are all serve as an intermediate step to control the expression of targeted genes.

The goal to evaluate gene expression can be achieved by several experimental transcriptome quantifying methods. Microarray based methods has been through more than

a decades of popularity (Schulze & Downward, 2001). For sequencing based methods, evaluation at transcript level expression can be done by RNA-seq (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008); methods capturing TSS include CAGE (Kodzius et al., 2006; Shiraki et al., 2003)and RNA-PET (Ruan et al., 2007). The experimental methods to quantify gene expression can be usually done with a lower price than profiling other epigenetic marks, due to the fact that the transcriptome is only a small portion of the whole genome on sequence length. Thus prediction methods are not proposed in order to perform *in silico* imputation for unmeasured data, instead, are focusing on studying the regulatory roles of different epigenetic marks on gene expression.

Initially there were efforts to study the patterns of DNA sequence at the regulatory regions of expressed genes in limited tissue context. Multiple sequence based prediction works are analyzed in an early work by Yuan et al., in which they reexam the sequence based prediction works, and proposed a method to select regulatory motif to predict gene expression in Yeast (Yuan, Guo, Shen, & Liu, 2007) .

In the meantime, histone modifications have been found to be predictive of gene expression. Histone modifications have many different types, depending on the location and the chemical group of the modification. This results in the diverse functional activities that they are able to initiate. Thus, it is necessary to treat different histone marks in different ways when modeling their regulatory roles. In the beginning, Yu et al. proposed to model the causal relationship between histone modifications and gene expression (Yu, Zhu, Zhou, Xue, & Han, 2008). Their Bayesian network not only reveals regulatory rules from histone modification to gene expression, but also inter-regulatory relationships among different types of histone marks. In a following study, Karlic et al. shows that only a small subset of

histone marks are needed to accurately predict gene expression levels, while the subset selection of histone marks is dependent on the GC content of the gene promoter regions (Karlić, Chung, Lasserre, Vlahovicek, & Vingron, 2010).

In addition to histone modifications, transcription factor binding sites (TFBS) are shown to be related to gene expression as well, which is not surprising due to the role of TFs. Ouyang et al. firstly propose to decompose groups of TFBS signals into sets of combinations (PCs) using PCA, then perform regression to model their relationships with gene expression (Ouyang, Zhou, & Wong, 2009). Since 2011, researcher started to combine multiple resources for the prediction task. Costa, et al. combines TFBS and histone marks to predict the expression levels of genes with low-CG promoters in diverse immune cells(Costa, Roider, do Rego, & de Carvalho, 2011). They point out that genes with low-CG promoters turn to express in tissue-specific manner. Cheng et al. followed the same strategy to predict gene expression in *C. elegans* and mouse ESCs, using SVM to accommodate model non-linearity(Cheng et al., 2011; Cheng & Gerstein, 2012). Current stage of prediction work still look for hint from histone marks. A recent deep learning predictive model used histone modification profile as the input features of a convolutional neural network and achieved good performance (Singh, Lanchantin, Robins, & Qi, 2016). New types of features have been added to predictive frameworks all the time. Park et al. used methylation levels as input for the first time in (Park & Nakai, 2011). This strategy has been extended to non-linear SVM model by Kapourani et al. in (Kapourani & Sanguinetti, 2016). Natarajan et al. introduced DNase hypersensitive sites (DHS) as the predicting feature when the ENCODE project release this type of cell-specific data, making the use of chromatin accessibility in the study of cell-specific epigenetic regulation in large

scale possible(Natarajan, Yardimci, Sheffield, & Frazer, 2012). In addition, DNA shape was also found to be predictive for TFBS and gene expression (Peng & Sinha, 2016).

Table 1.5 Prediction methods for gene expression

Publication	Features Used	Method	Target genes	Description
Yuan, et al. (2007), Plos Computational Biology	TF motifs	Naïve Bayes	Microarray, 2,587 Yeast genes	Revisit of sequence based expression prediction methods.
Yu, et al. (2008), Genome Research	Histone	Bayesian Network	Microarray , ~15,000 Human genes from multiple tissues	Predict regulatory network among histone and gene expression
Ouyang et al. (2009), PNAS	TFBS	Linear regession	RNA-seq, mouse ESCs	PCA on TF binding strength; use PCs as regressors
Karlic et al. (2010), PNAS	Histone	Linear regression	Microarray, human T-cell	Different combination of histone marks;
Costa, et al. (2011), BMC Bioinformatics	Histone + TFBS	Mixture of linear models	Microarray, human Th1, Th2, Th17 and iTreg cells	linear regression for each factor; then EM for estimating mixture
Park, et al. (2011), BMC Bioinformatics	12 TFBS, Methylation, Histone, CpG island	Linear regression	RNA-seq, mouse ESCs	Identified two classes of genes regulated by distinct combination of epigenetic marks

Cheng, et al. (2011), Genome Biology	Histone, TFBS,	SVM	RNA-seq, C. elegans and other species(modENCODE data)	Histone features are redundant; Positional contribution varies.
Natarajan, et al. (2012), Genome Research	DHS	Logistic regression + L_1 norm	Microarray, 19 human cell lines (from ENCODE)	Using DHS + TF motif to infer binding sites
Cheng, et al. (2012), NAR	Histone, TFBS	SVM	RNA-seq, mouse ESCs	TFBS and Histone shows distinct spatial patterns
Dong, et al. (2012), Genome Biology	Histone, DHS,	Random forest for binary prediction; regression to quantify expression	CAGE data, RNA-seq, RNA-PET(ENCODE)	Different subset of chromatin features are predictive for different types of RNA quantifying experiments
Kapourani, Sanguinetti. (2016) Bioinformatics	Methylation(RRBS)	SVM regression	RNA-seq, cell lines(K562, GM12878, H1-hESC)	Methylation profiles are predictive of gene expression across cell lines.
Singh, et al. (2016) Bioinformatics	Histone	CNN	RNA-seq (REMC)	Deep learning used
Peng and Sinha, (2016) NAR	DNA shape features, TF motifs	Random Forest	37 *Drosophila* genes	Only one using DNA shape features

1.7 Discussion

Here I presented the majority of predictive models in epigenetic studies. The pros and cons for computational methods are summarized based on this collection of methods. For the pros of predictive models, in-depth understanding of the regulatory cascade of genetic information is explicitly modeled in multiple scales and validated across many scenarios. It is also feasible to provide decision support for tests in practical applications, such as clinical settings. However, the drawbacks of in silico modeling should be noted as well. Firstly, the definition of gold standard is not unique, partly because of the fast development of experimental technologies. For the same predicting target, the definition of positive class may vary in both quantity and the genome coverage scale. Also the target itself can be highly heterogeneous and not well characterized, and limited by the resolution and accuracy of the experiment methods. Thus, the collection of data is a non-trivial task in order to fully saturate the feature space for all the potential sub-clusters of samples. Secondly, the rationale behind each prediction should be emphasized before exploring the correlation from data. Prediction tasks that using multiple marks requiring higher cost to predict lower-cost targets are impractical, and should not be considered. The economic reason is not the sole deal-breaker indeed, the biological logic is more important. For example, predicting gene expression using a full collection of chromatin profiles would be neither practical nor reasonable. Thirdly, although the correlation among features can be modeled, it is not sufficient to draw any conclusion on the causal relationships yet. One example of such inequity is the difficulty to transfer learned model or rules to a different context of cell other than the trained data. How well the model generalize the rules of transcription regulation is highly dependent on how representative the training data are.

Users need to be very careful when trying to make cross-cell, or cross-sample prediction, in order to make sure that the epigenome controlling rules can be transferred. Finally, the models are based on assumptions ignoring multiple other relevant genetic factors, such as the effect from genetic variants, modification on transcription, alternative splicing of transcripts, artifacts of techniques, etc. It is impossible to contain everything in one model. But, users should be aware to exclude the confounding factors to maximize the explain ability of proposed models.

By comparing models on the same topic in chronical order, there are interesting trend of model development on the same predicting topic. In general, the control flow of genetic information starts from DNA sequence, and ends at the expression of genes. At early stages, DNA sequence and features derived from it are the most common start point to build models. These studies are the initial exploring works, thus a well-developed software to repeat the same task is not common. Or, machine learning methods like SVM are widely used for high-dimension features based on k-mer sequence features. In later stages when the dynamic features can be profiled by new techniques, the sequence based models will generally lose popularity. Later models include more features, so the requirement for the complexity of models and the capability to handle heterogeneous features is essential in order to achieve better result. Models like random forest or boosting are popular choices. It is noteworthy that explicit modeling of experiment data plays a key role at all stages. A typical example is the characterization of DNase footprint shapes. Statistical models are powerful tools to quantitatively associate the observed sequencing data (read counts) to the interested biological events or characters (intensity, slope or other shape features of DNase

footprints). In the meantime, trading-off between model performance and model interpretability is also essential in the study of epigenetic regulation.

Reference

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. http://doi.org/10.1038/nbt.3300

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–61. http://doi.org/10.1038/nature12787

Angermueller, C., Lee, H. J., Reik, W., & Stegle, O. (2017). DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, *18*(1). http://doi.org/10.1186/s13059-017-1189-z

Arvey, A., Agius, P., Noble, W. S., & Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, *22*(9), 1723–1734. http://doi.org/10.1101/gr.127712.111

Barrera, L. A., Vedenko, A., Kurland, J. V., Rogers, J. M., Gisselbrecht, S. S., Rossin, E. J., … Bulyk, M. L. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, *351*(6280), 1450–1454. http://doi.org/10.1126/science.aad2257

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., … Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. http://doi.org/10.1016/j.cell.2007.05.009

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. http://doi.org/10.1038/nature07517

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74. http://doi.org/10.1038/nature11247

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., … Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), 1045–1048. http://doi.org/10.1038/nbt1010-1045

Bhasin, M., Zhang, H., Reinherz, E. L., & Reche, P. A. (2005). Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Letters*, *579*(20), 4302–4308. http://doi.org/10.1016/j.febslet.2005.07.002

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., … Pennacchio, L. A. (2010). ChIP-seq identification of weakly conserved heart enhancers. *Nature Genetics*. http://doi.org/10.1038/ng.650

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., … Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, *44*(2), 148–156. http://doi.org/10.1038/ng.1064

Brackley, C. A., Brown, J. M., Waithe, D., Babbs, C., Davies, J., Hughes, J. R., … Marenduzzo, D. (2016). Predicting the three-dimensional folding of cis -regulatory

regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biology*, 1–16. http://doi.org/10.1186/s13059-016-0909-0

Chen, X., Yu, B., Carriero, N., Silva, C., & Bonneau, R. (2017). Mocap: Large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Research*, *45*(8), 4315–4329. http://doi.org/10.1093/nar/gkx174

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., … Gerstein, M. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research*, *22*(9), 1658–1667. http://doi.org/10.1101/gr.136838.111

Cheng, C., & Gerstein, M. (2012). Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Research*, *40*(2), 553–568. http://doi.org/10.1093/nar/gkr752

Cheng, C., Yan, K.-K., Yip, K. Y., Rozowsky, J., Alexander, R., Shou, C., & Gerstein, M. (2011). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology*, *12*(2), R15. http://doi.org/10.1186/gb-2011-12-2-r15

Costa, I. G., Roider, H. G., do Rego, T. G., & de Carvalho, F. D. a T. (2011). Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics*, *12 Suppl 1*(Suppl 1), S29. http://doi.org/10.1186/1471-2105-12-S1-S29

Cotney, J., Leng, J., Oh, S., DeMare, L. E., Reilly, S. K., Gerstein, M. B., & Noonan, J.

P. (2012). Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Research*, *22*(6), 1069–1080. http://doi.org/10.1101/gr.129817.111

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., … Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, *107*(50), 21931–21936. http://doi.org/10.1073/pnas.1016071107

Cuellar-Partida, G., Buske, F. A., McLeay, R. C., Whitington, T., Noble, W. S., & Bailey, T. L. (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, *28*(1), 56–62. http://doi.org/10.1093/bioinformatics/btr614

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., … Zhang, M. Q. (2006). Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(28), 10713–6. http://doi.org/10.1073/pnas.0602949103

Erwin, G. D., Oksenberg, N., Truty, R. M., Kostka, D., Murphy, K. K., Ahituv, N., … Capra, J. A. (2014). Integrating Diverse Datasets Improves Developmental Enhancer Prediction, *10*(6). http://doi.org/10.1371/journal.pcbi.1003677

Fan, S., Huang, K., Ai, R., Wang, M., & Wang, W. (2016). Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. *Genomics*, *107*(4), 132–137. http://doi.org/10.1016/j.ygeno.2016.02.005

Fang, F., Fan, S., Zhang, X., & Zhang, M. Q. (2006). Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, *22*(18), 2204–2209.

http://doi.org/10.1093/bioinformatics/btl377

Firpi, H. A., Ucar, D., & Tan, K. (2010). Discover regulatory DNA elements using chromatin signatures and artificial neural network, *26*(13), 1579–1586. http://doi.org/10.1093/bioinformatics/btq248

Fortin, J.-P., & Hansen, K. D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, *16*(1), 180. http://doi.org/10.1186/s13059-015-0741-y

Ghandi, M., Lee, D., Mohammad-Noori, M., & Beer, M. A. (2014). Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology*, *10*(7). http://doi.org/10.1371/journal.pcbi.1003711

Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., … Natoli, G. (2010). Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. *Immunity*, *32*(3), 317–328. http://doi.org/10.1016/j.immuni.2010.02.008

Grunau, C. (2001). MethDB--a public database for DNA methylation data. *Nucleic Acids Research*, *29*(1), 270–274. http://doi.org/10.1093/nar/29.1.270

Gusmao, E. G., Allhoff, M., Zenke, M., & Costa, I. G. (2016). Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, *13*(4), 303–9. http://doi.org/10.1038/nmeth.3772

Gusmao, E. G., Dieterich, C., Zenke, M., & Costa, I. G. (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, *30*(22), 3143–3151. http://doi.org/10.1093/bioinformatics/btu519

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., … Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. *Science*, *320*(5872), 106–109. http://doi.org/10.1126/science.1150427

He, A., Kong, S. W., Ma, Q., & Pu, W. T. (2011). Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proceedings of the National Academy of Sciences*, *108*(14), 5632–5637. http://doi.org/10.1073/pnas.1016959108

He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., … Liu, X. S. (2010). Nucleosome dynamics define transcriptional enhancers. *Nature Genetics*, *42*(4), 343–7. http://doi.org/10.1038/ng.545

He, Y., Gorkin, D. U., Dickel, D. E., Nery, J. R., Castanon, R. G., Lee, A. Y., … Ecker, J. R. (2017). Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proceedings of the National Academy of Sciences*, 201618353. http://doi.org/10.1073/pnas.1618353114

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., … Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108–112. http://doi.org/10.1038/nature07829

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., … Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, *39*(3), 311–318. http://doi.org/10.1038/ng1966

Huang, J., Marco, E., Pinello, L., & Yuan, G.-C. (2015). Predicting chromatin organization using histone marks. *Genome Biology*, *16*(1), 162.

http://doi.org/10.1186/s13059-015-0740-z

Jankowski, A., Tiuryn, J., & Prabhakar, S. (2016). Romulus: Robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics*, *32*(16), 2419–2426. http://doi.org/10.1093/bioinformatics/btw209

Ji, H., Li, X., Wang, Q.-F., & Ning, Y. (2013). Correction for Kachar et al., High-resolution structure of hair-cell tip links. *Proceedings of the National Academy of Sciences*, *110*(19), 6789–6794. http://doi.org/10.1073/pnas.1311228110

Jia, C., & He, W. (2016). EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features Cangzhi Jia & Wenying He. *Nature Publishing Group*, (1), 1–7. http://doi.org/10.1038/srep38741

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., & Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions. *Nature Genetics*, *41*(8), 941–945. http://doi.org/10.1038/ng.409

Jung, S., Angarica, V. E., Andrade-Navarro, M. A., Buckley, N. J., & Del Sol, A. (2017). Prediction of Chromatin Accessibility in Gene-Regulatory Regions from Transcriptomics Data. *Scientific Reports*, *7*(1). http://doi.org/10.1038/s41598-017-04929-6

Kapourani, C. A., & Sanguinetti, G. (2016). Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, *32*(17), i405–i412. http://doi.org/10.1093/bioinformatics/btw432

Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National*

Academy of Sciences of the United States of America, *107*(7), 2926–31. http://doi.org/10.1073/pnas.0909344107

Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2014). DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Research*, *43*(1), e6–e6. http://doi.org/10.1093/nar/gku1058

Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaöz, U., Clelland, G. K., … Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research*, *17*(6), 691–707. http://doi.org/10.1101/gr.5704207

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., … Carninci, P. (2006). Cage: Cap analysis of gene expression. *Nature Methods*, *3*(3), 211. http://doi.org/10.1038/nmeth0306-211

Kuang, Z., Ji, Z., Boeke, J. D., & Ji, H. (2017). Dynamic motif occupancy (DynaMO) analysis identifies transcription factors and their binding sites driving dynamic biological processes. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkx905

Lee, D., Karchin, R., & Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Research*, *21*(12), 2167–2180. http://doi.org/10.1101/gr.121905.111

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., … Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322. http://doi.org/10.1038/nature08514

Liu, B., Fang, L., Long, R., Lan, X., & Chou, K. C. (2015). iEnhancer-2L: A two-layer

predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, *32*(3), 362–369. http://doi.org/10.1093/bioinformatics/btv604

Liu, F., Li, H., Ren, C., Bo, X., & Shu, W. (2016). PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *bioRxiv*, (March), 36129. http://doi.org/10.1101/036129

Liu, S., Zibetti, C., Wan, J., Wang, G., Blackshaw, S., & Qian, J. (2017). Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. *BMC Bioinformatics*, *18*(1). http://doi.org/10.1186/s12859-017-1769-7

Lu, Y., Qu, W., Shan, G., & Zhang, C. (2015). DELTA: A distal enhancer locating tool based on adaboost algorithm and shape features of chromatin modifications. *PLoS ONE*, *10*(6), 1–20. http://doi.org/10.1371/journal.pone.0130622

Ma, W., Yang, L., Rohs, R., & Noble, W. S. (2017). DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics*, *33*(19), 3003–3010. http://doi.org/10.1093/bioinformatics/btx336

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. http://doi.org/10.1038/nature03959

May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., … Visel, A. (2012). Large-scale discovery of enhancers from human heart tissue. *Nature Genetics*, *44*(1), 89–93. http://doi.org/10.1038/ng.1006

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E.

F., … Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*(9), 1527–1541. http://doi.org/10.1101/gr.091868.109

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, *33*(18), 5868–5877. http://doi.org/10.1093/nar/gki901

Miranda-saavedra, D. (2012). Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines, *40*(10). http://doi.org/10.1093/nar/gks149

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. http://doi.org/10.1038/nmeth.1226

Natarajan, A., Yardimci, G. G., Sheffield, N. C., & Frazer, K. a. (2012). Predicting cell-type – specific gene expression from regions of open chromatin the genome. *Genome Research*, 1711–1722. http://doi.org/10.1101/gr.135129.111

Ouyang, Z., Zhou, Q., & Wong, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(51), 21521–21526. http://doi.org/10.1073/pnas.0904863106

Paige, S. L., Thomas, S., Stoick-Cooper, C. L., Wang, H., Maves, L., Sandstrom, R., … Murry, C. E. (2012). A temporal chromatin signature in human embryonic stem cells

identifies regulators of cardiac development. *Cell*, *151*(1), 221–232. http://doi.org/10.1016/j.cell.2012.08.027

Park, S.-J., & Nakai, K. (2011). A regression analysis of gene expression in ES cells reveals two gene classes that are significantly different in epigenetic patterns. *BMC Bioinformatics*, *12 Suppl 1*(Suppl 1), S50. http://doi.org/10.1186/1471-2105-12-S1-S50

Peng, P., & Sinha, S. (2016). Quantitative modeling of gene expression using DNA shape features of binding sites. *Nucleic Acids Research*, *44*(13), gkw446. http://doi.org/10.1093/nar/gkw446

Pique-regi, R., Degner, J. F., Pai, A. A., Boyle, A. P., Song, L., Lee, B., … Pritchard, J. K. (2011). sequence and chromatin accessibility data Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data, 447–455. http://doi.org/10.1101/gr.112623.110

Qin, Z., Li, B., Conneely, K. N., Wu, H., Hu, M., Ayyala, D., … Lin, S. (2016). Statistical Challenges in Analyzing Methylation and Long-Range Chromosomal Interaction Data. Statistics in Biosciences, 8(2), 284-309. http://doi.org/10.1007/s12561-016-9145-0

Quach, B., & Furey, T. S. (2017). DeFCoM: Analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics*, *33*(7), 956–963. http://doi.org/10.1093/bioinformatics/btw740

Quang, D., & Xie, X. (2017). FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *bioRxiv*, 151274. http://doi.org/10.1101/151274

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279–285. http://doi.org/10.1038/nature09692

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., … Ren, B. (2013). RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Computational Biology*, *9*(3). http://doi.org/10.1371/journal.pcbi.1002968

Ramsey, S. A., Knijnenburg, T. A., Kennedy, K. A., Zak, D. E., Gilchrist, M., Gold, E. S., … Shmulevich, I. (2010). Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, *26*(17), 2071–2075. http://doi.org/10.1093/bioinformatics/btq405

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–329. http://doi.org/10.1038/nature14248

Rollins, R. A., Haghighi, F., Edwards, J. R., Das, R., Zhang, M. Q., Ju, J., & Bestor, T. H. (2006). Large-scale structure of genomic methylation patterns. *Genome Research*, *16*(2), 157–163. http://doi.org/10.1101/gr.4362006

Ruan, Y., Hong, S. O., Siew, W. C., Kuo, P. C., Xiao, D. Z., Srinivasan, K. G., … Wei, C. L. (2007). Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Research*, *17*(6), 828–838. http://doi.org/10.1101/gr.6018607

Sandelin, A. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, *32*(90001), 91D–94.

http://doi.org/10.1093/nar/gkh012

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., … Zhao, K. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, *132*(5), 887–898. http://doi.org/10.1016/j.cell.2008.02.022

Schulze, A., & Downward, J. (2001). Navigating gene expression using microarrays - A technology review. *Nature Cell Biology*. http://doi.org/10.1038/35087138

Shendure, J. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, *309*(5741), 1728–1732. http://doi.org/10.1126/science.1117389

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., Van Hoff, J. P., … Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, *32*(2), 171–178. http://doi.org/10.1038/nbt.2798

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., … Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, *100*(26), 15776–15781. http://doi.org/10.1073/pnas.2136655100

Singh, R., Lanchantin, J., Robins, G., & Qi, Y. (2016). DeepChrome: Deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, *32*(17), i639–i648. http://doi.org/10.1093/bioinformatics/btw427

Sung, M. H., Guertin, M. J., Baek, S., & Hager, G. L. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Molecular Cell*, *56*(2), 275–285. http://doi.org/10.1016/j.molcel.2014.08.016

Taher, L., Narlikar, L., & Ovcharenko, I. (2012). Clare: Cracking the LAnguage of regulatory elements. *Bioinformatics*, *28*(4), 581–583. http://doi.org/10.1093/bioinformatics/btr704

Talebzadeh, M., & Zare-Mirakabad, F. (2014). Transcription factor binding sites prediction based on modified nucleosomes. *PLoS ONE*, *9*(2). http://doi.org/10.1371/journal.pone.0089226

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82. http://doi.org/10.1038/nature11232

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., … Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–858. http://doi.org/10.1038/nature07730

Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., … Bruneau, B. G. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, *151*(1), 206–220. http://doi.org/10.1016/j.cell.2012.07.035

Wang, J., Zhuang, J., Iyer, S., Lin, X. Y., Greven, M. C., Kim, B. H., … Weng, Z. (2013). Factorbook.org: A Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Research*, *41*(D1). http://doi.org/10.1093/nar/gks1221

Wang, Y., Liu, T., Xu, D., Shi, H., Zhang, C., Mo, Y.-Y., & Wang, Z. (2016). Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. *Scientific Reports*, *6*(January), 19598.

http://doi.org/10.1038/srep19598

Whitaker, J. W., Chen, Z., & Wang, W. (n.d.). Predicting the human epigenome from dnA motifs. http://doi.org/10.1038/nmeth.3065

Whitington, T., Perkins, A. C., & Bailey, T. L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Research*, *37*(1), 14–25. http://doi.org/10.1093/nar/gkn866

Won, K. J., Ren, B., & Wang, W. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, *11*(1), R7. http://doi.org/10.1186/gb-2010-11-1-r7

Xu, T., Li, B., Zhao, M., Szulwach, K. E., Street, R. C., Lin, L., … Qin, Z. S. (2015). Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Research*, 1–10. http://doi.org/10.1093/nar/gkv151

Yardimci, G. G., Frank, C. L., Crawford, G. E., & Ohler, U. (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Research*, *42*(19), 11865–11878. http://doi.org/10.1093/nar/gku810

Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., … Gerstein, M. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, *13*(9). http://doi.org/10.1186/gb-2012-13-9-r48

Yu, H., Zhu, S., Zhou, B., Xue, H., & Han, J. D. J. (2008). Inferring causal relationships among different histone modifications and gene expression. *Genome Research*,

18(8), 1314–1324. http://doi.org/10.1101/gr.073080.107

Yuan, Y., Guo, L., Shen, L., & Liu, J. S. (2007). Predicting gene expression from sequence: A reexamination. *PLoS Computational Biology*, *3*(11), 2391–2397. http://doi.org/10.1371/journal.pcbi.0030243

Zeng, H., & Gifford, D. K. (2017). Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Research*, *45*(11), e99. http://doi.org/10.1093/nar/gkx177

Zentner, G. E., Tesar, P. J., & Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Research*, *21*(8), 1273–1283. http://doi.org/10.1101/gr.122382.111

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., & Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology*, *16*(1), 14. http://doi.org/10.1186/s13059-015-0581-9

Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., … Wang, W. (2016). Constructing 3D interaction maps from 1D epigenomes. *Nature Communications*, *7*, 10812. http://doi.org/10.1038/ncomms10812

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., & Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, *462*(7269), 65–70. http://doi.org/10.1038/nature08531

Zou, L. S., Erdos, M. R., Taylor, D. L., Chines, P. S., Varshney, A., Institute, T. M. G., … Didion, J. P. (2018). BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *bioRxiv*, 207506. http://doi.org/10.1101/207506

Chapter 2
Base-resolution methylation patterns accurately predict transcription factor bindings *in vivo*

2.1 Introduction

A fundamental goal of functional genomic research is to understand gene regulation. Gene expression can be controlled by epigenetic mechanisms via the coordinated binding of transcription factors (TFs), histone modifications, and DNA methylation (Cooper & Hausman, 2013). An important first step toward deciphering the complexities of gene regulatory networks is detecting the activities of functional elements, such as TF binding sites in the genome.

Advances in high-throughput sequencing technologies such as ChIP-seq (Barski et al., 2007; Johnson, Mortazavi, Myers, & Wold, 2007; Robertson et al., 2007) and ChIP-exo (Rhee & Pugh, 2011) allow the comprehensive genome-wide profiling of protein-DNA binding sites. In recent years, enormous efforts have been made to map TF binding sites under different biological contexts; for example, by consortiums like ENCODE (Consortium et al., 2012) and modENCODE (Celniker et al., 2009). In spite of the successes, the application of ChIP-seq is still limited by the availability of high-quality antibodies and a requirement for fresh cells/tissues. The multitude of distinct proteins makes genome-wide profiling for all of them labor-intensive and costly. Furthermore, individual profiling of TF binding is a challenge in clinical settings because the amount of biological materials available is often limited. For these reasons, developing in silico approaches to predict in vivo TF binding sites that do not rely on ChIP-seq is desirable.

Traditionally, DNA sequence motifs have been used to predict TF binding (Stormo, 2000; Tompa et al., 2005). However, such an approach only works well for proteins with binding motifs that are highly specific. For proteins with weak binding motif patterns, the predictions suffer low specificity. In addition, the DNA motif is insufficient to determine whether a TF will bind to DNA in vivo, which means cell type-specific binding cannot be determined; additional information is needed to make that prediction. Recent studies revealed that TF binding is associated with nucleosome positions (He et al., 2010), histone marks (Heintzman et al., 2007; Robertson et al., 2007), and hypersensitivity to cleavage by DNase I (Bernat et al., 2006; Hesselberth et al., 2009). Based on these findings, a number of statistical methods and software tools have been developed to integrate motif information with other data types and genome annotations to achieve better prediction results (Arvey, Agius, Noble, & Leslie, 2012; Cuellar-Partida et al., 2012; He et al., 2010; Ji, Li, Wang, & Ning, 2013; Pique-Regi et al., 2011; Rajagopal et al., 2013; Ramsey et al., 2010; Won, Ren, & Wang, 2010). All these methods use histone or DNase I data, as well as the genome annotations and DNA motifs for prediction. One of the common limitations is that the histone modification or DNase I hypersensitivity studies require large amounts of fresh starting material (at least from 10^6 cells). This makes the existing prediction methods practically inapplicable to clinical samples.

DNA methylation is an important epigenetic modification with essential roles in many biological processes (Klose & Bird, 2006; Suzuki & Bird, 2008). Methylation of cytosine at carbon five (5-methylcytosine, or 5mC) regulates gene expression, determines cell development, and affects numerous disease pathogeneses (Jones, 2012; Klose & Bird, 2006). Exploiting next-generation sequencing technologies, a powerful experimental assay

called bisulfite sequencing (BS-seq) was developed that measures DNA methylation at base resolution genome-wide (Lister et al., 2013; Lister et al., 2008; Lister et al., 2009). The experiment starts by treating DNA molecules with sodium, which induces deamination and conversion of unmethylated cytosine to uracil, while methylated cytosine is protected by the methyl group and remains unchanged. The uracil will be amplified as thymine during amplification. The bisulfite-treated and PCR-amplified DNA segments then go through high-throughput sequencing. After alignment and preprocessing, BS-seq data can be analyzed by counting the number of sequencing reads for each CpG site where either a thymine or a cytosine is observed. The count of thymine represents the number of sequenced DNA strands that are unmethylated, and the count of cytosine represents the number of DNA strands that are methylated at this CpG site.

5mC is known to interfere with DNA-protein interactions, thereby directing transcriptional states (Hu et al., 2013). For example, a recent publication reported that 5mC is strongly correlated with TF binding, where the binding sites are usually hypomethylated (Stadler et al., 2011). Regulation of DNA-protein interactions can occur either through affinity of methyl-CpG-binding proteins for 5mC, or through the refractory effects of 5mC on some DNA-protein interactions. The latter is known to directly influence binding of a number of TFs, such as CTCF (Yu et al., 2012). Furthermore, more recent observations have implicated the iterative oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcystosine (5caC) in pathways that serve to offset 5mC levels and facilitate TF binding (Song et al., 2013). All these findings indicate that DNA methylation levels offer clues as to whether TF binding occurred at a particular locus, which may be exploited as an alternative to the DNase I or histone data for the purpose of

predicting TF binding in vivo. This is important because DNA methylation profiles are more stable and much easier to obtain than DNase I profiles in a clinical setting.

To investigate the viability of this hypothesis, we aligned profiles of DNase I, methylation, and TF binding obtained by DNase-seq, BS-seq, and ChIP-seq, respectively, in which the ChIP-seq data are used as the gold standard for TF binding. Visual inspection showed there are good concordances between ChIP-seq peaks, DNase I peaks, and methylation "dips." As an example, Figure 2.1 shows one CTCF binding sites in H1-hESC, which is located at the transcription start site (TSS) of a protein-coding gene. One can clearly see that at the TF binding sites (indicated by ChIP-seq peaks), the DNase-seq data indicates enrichment of DNase I hypersensitivity sites. At the same locations, the DNA methylation levels are altered and show strong hypomethylation.

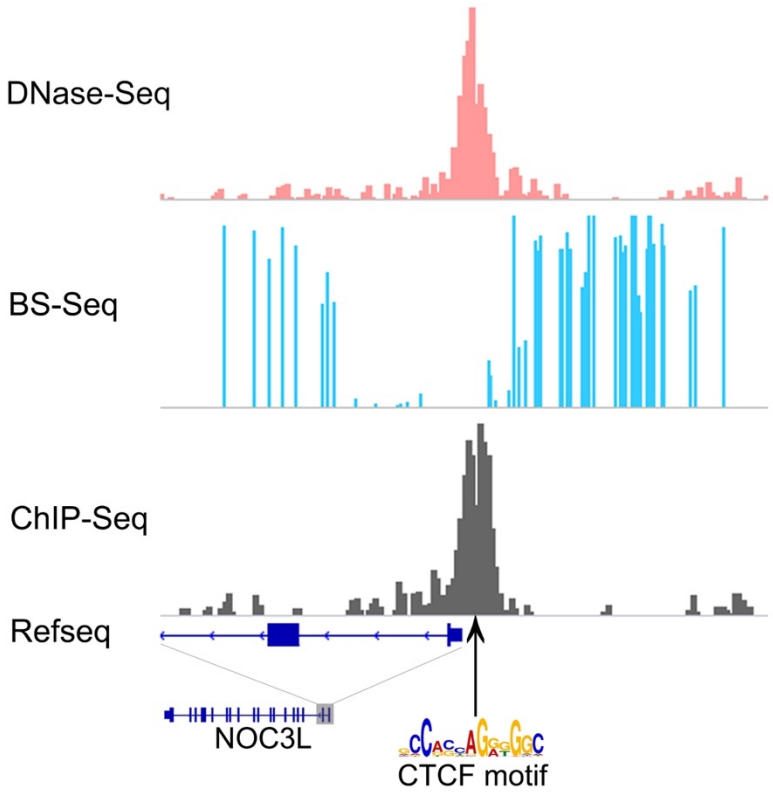

Figure 2.1 Concordance among epigenetic profiles

Concordances between ChIP-seq, DNase I and methylation on a genomic locus. A comparative view of DNase I- hypersensitive, methylation (5mC), and ChIP-seq profiles on a genomic locus on chromosome 10. Good concordances are shown between ChIP-seq peaks (used as the gold standard for TF binding), DNase I peaks, and methylation "dips". All data shown are from the H1-hESC cell line.

For a more comprehensive view of the methylation profiles around TFBSs, we explored whole-genome BS-seq data from two human cell lines (embryonic stem cell H1-hESC and fibroblast IMR90) and one mouse cell line (embryonic stem cell mESC). We calculated the average methylation levels of three types of methylation: CG methylation (5mC), CG

hydroxymethylation (5hmC), and non-CG (CH) methylation around putative TF binding sites (motif sites covered by a ChIP-seq peak) and compared these levels to those from non-TF binding sites (motif site not covered by any ChIP-seq peak). A "meta-gene" style plot is shown in Figure 2.2. From these plots, we make three important observations. First, there are differences in the methylation levels between actual TF binding sites and random regions, with 5mC patterns showing the most pronounced difference. Second, the methylation profiles are distinct for different TFs. Third, the methylation patterns for the same TF are similar across cell types. Taken together, these findings indicate that methylation profiles, similar to the DNase-seq data, can be used to distinguish TF binding sites from the genomic background. Despite the empirical evidence connecting methylation level variation and TF binding, how to develop a rigorous statistical approach to quantify the methylation profiles around TF binding sites is non-trivial. Another key question is how to integrate methylation information along with DNA sequence motif and other genomic features in a coherent framework to predict TF binding in vivo.

Motivated by these findings, we developed a novel computational approach to predict TF binding. Our method, named Methylphet, is a supervised learning strategy that is able to combine methylation profiles and multiple genomic features to make TF binding predictions. Using ChIP-seq data as surrogates for putative TF binding, we show that Methylphet achieves higher accuracy than prediction method using motif score alone or DNase I profiles. Compared with histone ChIP-seq or DNase-seq, BS-seq can be accomplished using very little material (nanograms of genomic DNA) with highly sensitive bisulfite conversion-based methods, making a prediction method based on BS-seq data a

good alternative means for inferring gene regulatory mechanisms from samples in which ChIP-seq and DNase I hypersensitivity studies are not feasible.

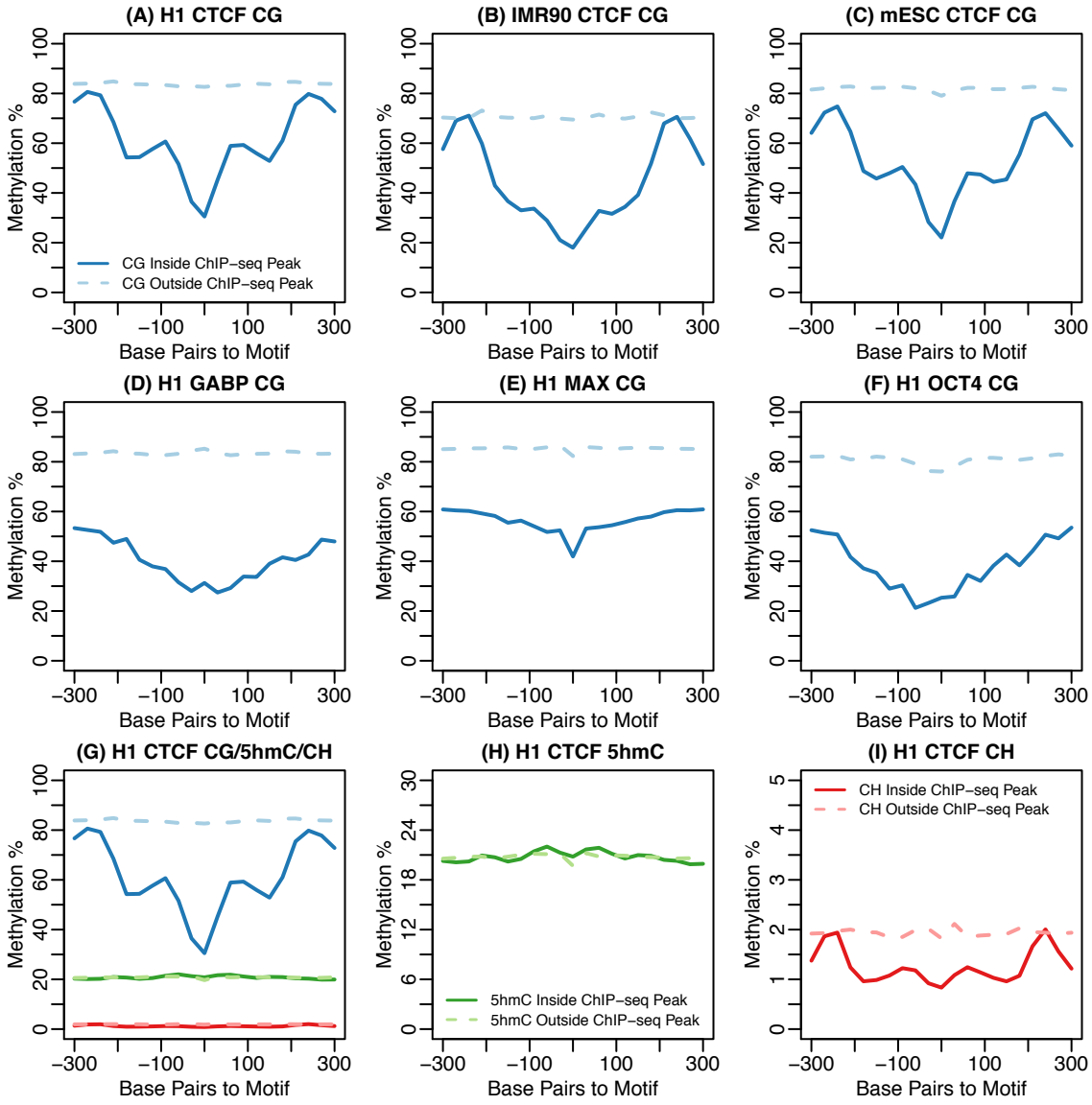

Figure 2.2 Methylation profiles from different cell lines/TFs/methylation type. Methylation patterns around binding sites of several TFs from different cell lines. Curves represent average methylation (CG/5hmC/CH) levels around ChIP-seq peaks (solid lines) and motif sites without ChIP-seq peaks (dashed lines). A-C: CG methylation profiles for CTCF from different cell lines (H1-hESC, IMR90, and mESC). D-F: CG methylation

profiles for several TFs (GABP/MAX/OCT4) from H1-hESC. G-I: Different types of methylation profiles (CG/5hmC/CH) for CTCF from H1-hESC.

2.2 Material and Methods

2.2.1 Description of the Methylphet method

The workflow of Methylphet is illustrated in Figure 2.3. The method consists of candidate site selections, a training module and a testing module. The first step is to identify candidate binding sites by genome-wide motif scan using motif PWMs in both training and testing data. The detailed procedure of selecting candidate sites is provided in the Data and Processing section. Then a predictive model is constructed in the training module, and then the model is put to work for TF binding prediction in the testing module. We use Random Forest (RF) (Breiman, 2001) to build the predictive model. RF is an ensemble learning method for classification that recently became popular in genomics because of its flexibility, efficiency, and ability to avoid over-fitting. Moreover, RF provides importance measurements for all predictors, which are key to deciding whether to remove an unrelated predictor or add a new promising one.

The required inputs for the training module include the ChIP-seq peak locations (as the gold standard), a set of whole-genome BS-seq data, and other static genomic features, such as DNA motif and evolutionary conservation scores. Optionally, 5-hydroxymethylcytosine (5hmC) data from Tet-assisted BS-seq (TAB-seq) (Yu et al., 2012) can also be included. The training module contains two steps: the construction of a methylation model and a RF model respectively. Motif information is not used in the methylation model training step,

but used in constructing the RF model. With the candidates available, we first identify the putative binding sites (those inside a ChIP-seq peak) from all candidate regions using the gold standard ChIP-seq data. Next in the estimation of methylation models, we characterize the methylation count data in a genomic window around the true TF binding sites as a series of beta-binomial distributions (details are provided in the next section). Then the same procedure is applied to candidate regions without TF binding. At the end of this step, we obtain two sets of beta-binomial distributions for the methylation profiles from TF binding and background regions. For example, means of the beta-binomial distributions represent the shapes of the methylation levels around TF binding or background regions (as shown in Figure 2.2). Based on the estimated distributions, for each candidate region we compute multiple "methylation scores," which are defined as the likelihood ratios of the site being a true binding site versus being the background. The methylation scores include 5mC scores, CH methylation scores, and 5hmC scores if TAB-seq data are available. Next for the training of the RF model, in addition to methylation scores, we also include genomic features, such as motif scores, conservation scores, and distance to TSS. Subsequently, the methylation and RF models produced from the training module are employed for prediction. It is important to note that a different predictive model is constructed for each TF due to the TF specificity of the methylation profiles.

2.2.2 Methylation models

We used the following model to characterize methylation (including 5mC, 5hmC, and CH methylation) patterns at a genomic region. Given a candidate site, we treat the motif site as a window, and then add ten 30bp window to each side. The methylation profiles in these 21 windows are used to capture methylation patterns for TF binding sites and backgrounds.

We choose 30bp as the window size for the sake of balancing the needs of model parameter estimation accuracy and spatial resolution of the methylation profile. To gauge the impact of the window size selection, we repeat the whole analysis procedure using window size of 20bp and compare the two sets of results. We found that the two sets of results are very similar and the 30-bp results are slightly better overall. In the software implementation of Methylphet, we provide option for user to specify the window size.

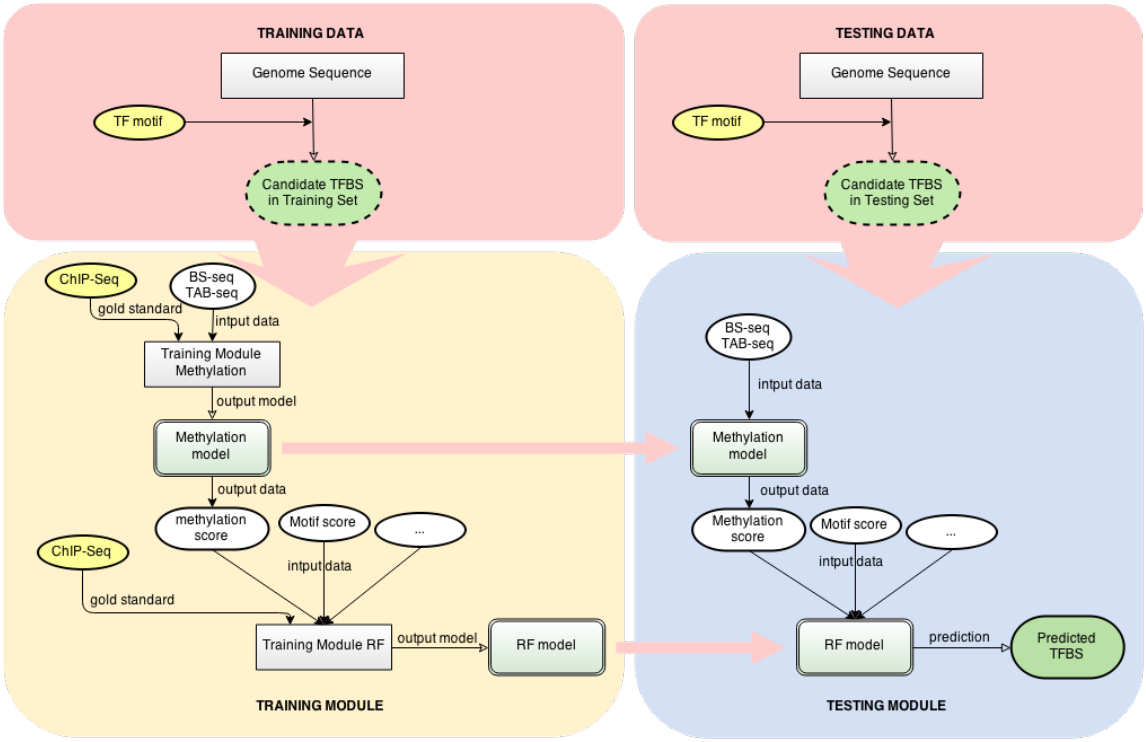

Figure 2.3 A flow chart for Methylphet method.

Inside each window, if there was at least one CG dinucleotide covered by at least one read (either methylated or not), we recorded the total number of methylated and unmethylated reads. Assume there are n candidate sites. In the jth window ($j = 1,2,\ldots,21$) of the ith

candidate site ($i = 1,2, \dots n$), we used x_{ij} and y_{ij} to denote the number of methylated and unmethylated reads and let $n_{ij} = x_{ij} + y_{ij}$.

Similar to in (Feng, Conneely, & Wu, 2014), we used a beta-binomial compound distribution to model the count data from BS-seq. The counts, given underlying "true" methylation levels, are assumed to follow a binomial distribution:

$$x_{ij}|n_{ij}, p_j \sim Binom(n_{ij}, p_j), i = 1,2, \dots n; j = 1,2, \dots 21.$$

The methylation levels p_j's are assumed to follow a beta distribution, but with different parameters at TF binding sites and background. When there is no TF binding at a candidate site, the methylation levels from all 21 windows are assumed to be identical and similar to those from the genomic background (close to fully-methylated). Thus, we assume p_j's follow the same beta distribution. For candidate sites that are bound by TFs, we found (Figure 2.2) that the methylation levels are different at different windows, e.g., methylation levels dip toward the motif site from both directions. Therefore, we assume that each p_j follows a different beta distribution. Defining indicator z_i to denote binding ($z_i = 1$) or not ($z_i = 0$) for candidate site i, we have

$$p_j|z_i = 1 \sim Beta(\alpha_j, \beta_j), \qquad p_j|z_i = 0 \sim Beta(\alpha', \beta').$$

For quality control purposes, we removed all windows with less than five total reads and candidates within CpG islands from the training set and used the method of moment (MOM) to estimate parameters α_j, β_j and α', β'. With parameters estimated at each motif site, we calculated the likelihood ratio comparing the two methylation patterns (TF binding or no binding) as methylation score λ_i for the ith candidate sites in test data:

$$\lambda_i = \sum_{j:\#\ of\ CG>0}^{m} \log\left(\frac{p(x_{ij}|n_{ij}, z_i = 1)}{p(x_{ij}|n_{ij}, z_i = 0)}\right)$$

Higher methylation scores indicated stronger evidence for a candidate site to have TF binding. The same procedure was applied to obtain CH methylation scores, as well as 5hmC scores if whole-genome TAB-seq data were available.

2.2.3 Other genomic features

Other genomic information used in the predicting model included: sequence conservation, distance to TSS, overlap with repetitive region, and other genomic features. Conservation scores were downloaded from the UCSC genome browser, hg18 phastCons44way table. Repeat masker, which marks the repetitive regions, was downloaded from the UCSC genome browser. We also calculated the nearest distance between candidate binding sites. For other genomic features, we used a binary indicator (0 or 1) to show if the motif overlapped with: TSS, TES, exons, introns, or CpG islands, and the distance to TSS. All the genomic Feature annotations were calculated using R and Bioconductor.

2.2.4 Prediction

Several supervised learning approaches were investigated, and RF performed the most accurately and robustly among all the approaches. Hence, results were demonstrated using RF, which was achieved with R package randomForest (Liaw & Wiener, 2002).

In the RF, a binary classification model was trained using methylation score together with genomic features. In each trained model, the importance of input features was assessed using Gini gain importance. The number of trees used in the model was determined by the stability of out-of-bag error. The predicting result is represented by the probability of getting a vote from the randomly generated classification tree for each class. The predicting performance was evaluated using the ChIP-seq peaks as the gold standard. ROC based on

the class probability and the gold standard was computed to show the overall predicting performance of our method.

2.2.5 Data and processing

5mC data from bisulfite sequencing (BS-seq) studies. The BS-seq data from human embryonic stem-cell (hESC) lines H1-hESC and IMR90 were downloaded from Gene Expression Omnibus (GEO) with ID GSE16256 (Lister et al., 2009). The 5mC BS-seq data from the mouse embryonic stem-cell (mESC) line was downloaded from GEO with ID GSE30202 (Stadler et al., 2011). The 5mC BS-seq data from the mouse dentate gyrus (DG) cells was downloaded from GEO with ID GSM1263221 (Guo et al., 2014)

Bisulfite-seq paired-end read processing and methylation calling. Paired-end reads were first pre-processed to remove adapter sequences, as well as low-quality sequence on both the 3' and 5' ends using Trimmomatic 0.20 (Bolger, Lohse, & Usadel, 2014), with the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. This was followed by in silico conversion of each C to T (Read 1) and each G to A (Read 2). Preprocessed reads were then aligned to both C-to-T and G-to-A converted chromosomes that were computationally derived from NCBI mm9 genomic sequence using Bowtie 0.12.9 (Langmead, Trapnell, Pop, & Salzberg, 2009) (-m 1 -l 30 -n 0 -e 90 -X 550). Reads mapping to both genomes were discarded and non-aligned pairs were reprocessed as single-end data using the same alignment parameters. For both paired-end and single-end alignments, only uniquely mapping reads were retained, and PCR duplicates were removed using MarkDuplicates (Picard Tools 1.82). To avoid counting reference positions covered by overlapping paired-end reads, overlapping regions were clipped, keeping the region of the overlap with higher quality. The original computationally

converted C's and G's were reverted, and for each reference cytosine position the number of C reads and T reads were counted using SAMTools mpileup. We kept the number of 5mC reads as well as total read coverage at each CG dinucleotide where 5mC is present for the processed data.

5hmC data. The base-resolution maps of 5hmC in human and mouse ES cells were generated previously (Yu et al., 2012). We used the same procedure described above to process 5hmC data and call methylation.

DNase data. The DNase I cutting sites were derived from the ENCODE dataset. We downloaded the Human H1-hESC DNase sequencing alignment files from ENCODE Crawford-Duke chromatin Map dataset via ENCODE Data Coordination Center (DCC); we downloaded the mouse mESC DNase sequencing alignment files from ENCODE Uw Dnase dataset from ENCODE DCC; DNase-seq alignment files for the IMR90 cell line were first downloaded from ENCODE Duke OpenChromDnase dataset on ENCODE DCC, and then converted to the hg18 coordinate system using liftover (Hinrichs et al., 2006).

ChIP-seq data. The ChIP-seq mapping results for all the TFs in H1-hESC cells, IMR90 cells, and CTCF in mouse mESC Cells were downloaded from UCSC ENCODE collection (Wang et al., 2012). We performed the ChIP-seq experiment on mouse mESC for OCT4.

ChIP-seq experiment. ChIP-seq experiments were performed following the protocol from the laboratory of Richard M. Myers (http://myers.hudsonalpha.org/documents/Myers%20Lab%20ChIP-seq%20Protocol%20v041610.pdf). Briefly, 2×10^{7} mouse ES cells were cross-linked with 1% formaldehyde at 25°C for 10 min and sonicated to generate chromatin fragments of

100–500 bp. Chromatin fragments from 2×10^7 cells were immunoprecipitated using OCT4 antibody (Abcam ab8895). ChIP-seq library construction and Illumina sequencing were performed following the manufacturer's instructions.

ChIP-seq data processing. For bam files that listed genomic coordinates in hg19, we first converted them to genomic coordinates in hg18 using liftover (Hinrichs et al., 2006). We next used HPeak (Qin et al., 2010) for peaking calling. Peak intersections of biological replicates are retained and employed for model training to maintain enhanced ChIP signal strength.

Selection of candidate regions. The candidate TF binding regions are selected based on sequence motif scores. The position-specific weight matrices (PWM) for CTCF, MAX, SIX5, USF1, BCL11A, EGR1, NANOG, RAD21, RFX5, SRF, USF2, GABP, NRSF, YY1, CJUN, JUND , OCT4, and TCF12 were downloaded from JASPAR (Mathelier et al., 2014) and factorbook (Wang et al., 2012). We used PWM matching functions from Bioconductor package "Biostrings" to scan the entire genome to identify candidate sites for TFBS. The cutoff for candidate sites leaves between 200,000 and 600,000 candidate sites for most TFs.

2.2.6 Data Access

ChIP-Seq data in mESC have been submitted to GEO (GEO accession number GSE65093).

2.3 Result

2.3.1 TF binding prediction results

We conducted extensive real data analyses to evaluate the performance of Methylphet. In total, we performed TF binding prediction of 19 TFs for human embryonic stem cell line

H1-hESC and five TFs for human fibroblast cell line IMR90. We randomly split candidate sites into equal-sized training set and testing set. Prediction performance is evaluated on the testing set only. We compared the prediction performance of Methylphet with CENTIPEDE (Pique-Regi et al., 2011), which is a widely used unsupervised method using DNase-seq data to predict TFBS. CENTIPEDE takes other genomic features to construct the prior for TF binding prediction. For a fair comparison, we fed the same set of non-cell-type-specific, static genomic features, such as conservation score, motif score, etc. to both CENTIPEDE and Methylphet. Besides those, DNase data were used in CENTIPEDE, and methylation data were used in Methylphet. We did not include methylation data for CENTIPEDE, nor did we include DNase data for Methylphet. We also compared with predictions using sequence motif only (candidate regions are ranked by their motif scores). Receiver Operation Characteristic (ROC) curves are used to represent the overall predicting performance of each method (Figure 2.4-5). Considering that a majority the candidate sites are negative for some TFs, we also generate precision-recall curves for performance evaluation.

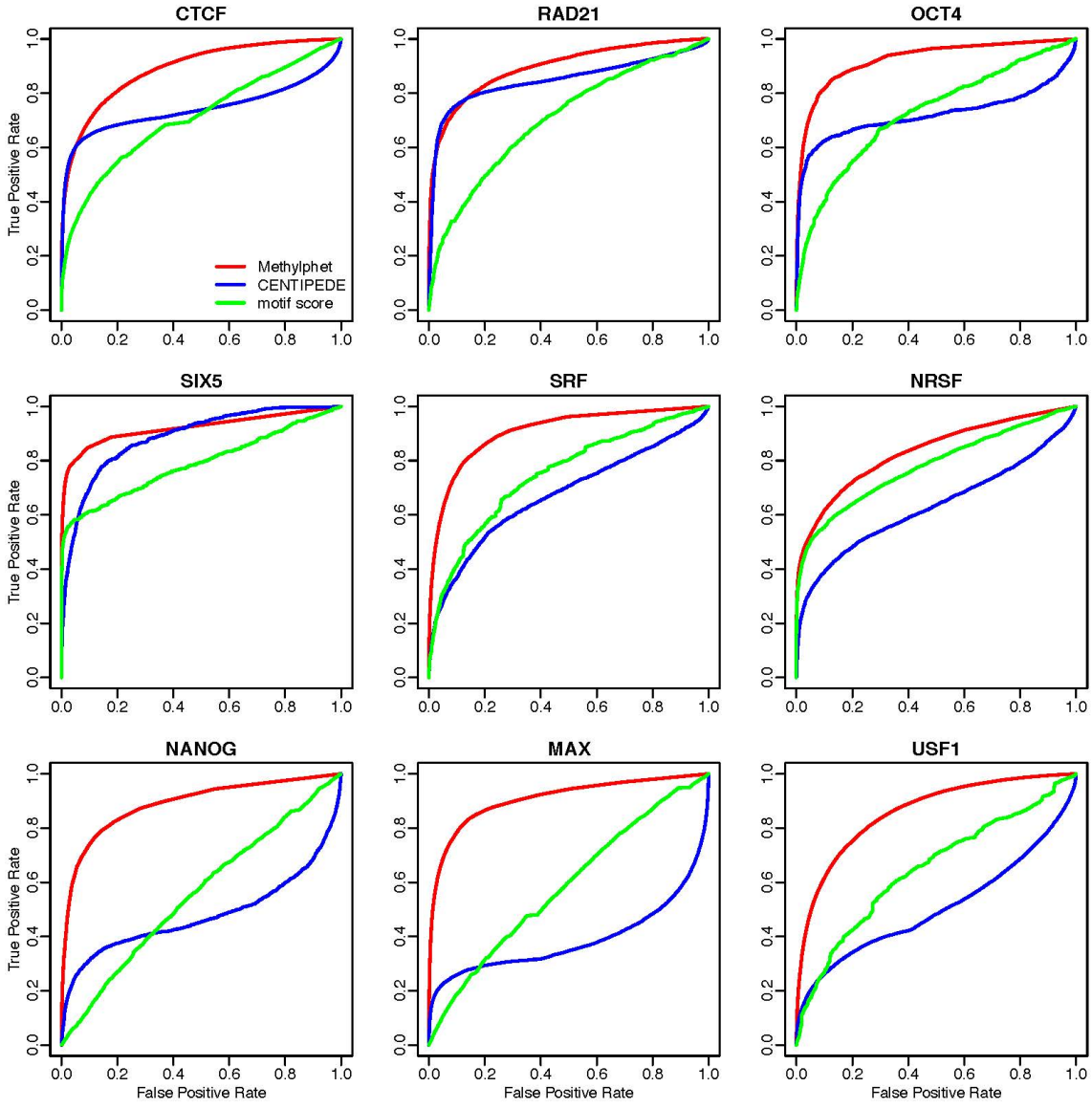

Figure 2.4 TF binding predition results for H1-hESC cell line.

ROC curves for Methylphet, CENTIPEDE, and motif score are shown by red, blue, green lines, respectively. Prediction results were generated by randomly splitting the dataset into training set and testing set of equal size for the H1-hESC cell line. Methylphet robustly provides better predicting performance for different transcription factors.

Figure 2.4 shows the ROC curves for nine different TFs in H1-hESC. These extensive real data analyses show that Methylphet robustly outperforms CENTIPEDE and the motif-score-only method. To be more specific, Methylphet outperforms motif-score-only for all TFs in all cell lines. This is expected since Methylphet effectively combines motif score and information from methylation profiles and other genomic features. Compared with CENTIPEDE, which also considers motif score and other genomic features, Methylphet also outperforms significantly in all TFs. Although CENTIPEDE and Methylphet perform similarly for CTCF and RAD21 when the false-positive rate (FPR) is small (less than 0.1), Methylphet outperforms CENTIPEDE after the FPR is higher than 0.1. In order to demonstrate the robustness of the performance, we repeat the testing for ten times. The boxplots of the ten area under the ROC curves values show that Methylphet robustly outperforms motif score and CENTIPEDE.

In addition to the difference in information sources, the underlying strategy of Methylphet, which is an ensemble learning approach, is also different from that of CENTIPEDE, which is a mixture model type of approach. It is of interest to find out whether the source of data, or the underlying method is the major contributor of the performance improvement of Methylphet. In order to answer the above question, we replaced the methylation scores with the DNase scores obtained from CENTIPEDE in the RF of Methylphet and compare that to Methylphet as well as CENTIPEDE. Our comparison results between RF with methylation data vs. RF with DNase data seem to suggest that both data source (methylation data vs. DNase data) and method used (RF vs. mixture model) contribute to the performance improvement of Methylphet over CENTIPEDE. However it is also possible that the statistical model, not the data source used, made the difference. Therefore,

an alternative model for DNase with RF could change, and potentially improve the predicting ability of DNase data.

The advantages of RF, an ensemble learning approach, over the mixture model type of approach adopted by CENTIPEDE can be attributed to two factors. First, due to the high variability among TFs, a supervised learning approach like RF is more robust. On the other hand, an unsupervised mixture model approach may fail in adverse situations. As an example, the EM algorithm (Dempster, Laird, & Rubin, 1977) fails to converge when the proportion of true positives in the candidate sites is low, which often occurs for TFs with shorter motifs and fewer putative binding sites. Second, RF does not assume independence among the predictors, as does CENTIPEDE. Our experience with CENTIPEDE is that the final results are often dominated by the DNase-seq data. Since the genomic features are diverse and many of them are highly correlated, an RF model can better use the integrated information from predictors.

2.3.2 Cross-sample TF binding prediction results

We further tested the predictive accuracy when training and testing data were from different samples. Our approach will be most attractive if the model trained in one cell type can produce robust prediction in a different cell type or sample; from Figure 2.2, this seems plausible since we saw that the methylation pattern is consistent across cell lines for the same TF.

To verify this, we conducted tests in which we trained the Methylphet model using data from IMR90 and mESC cell lines, and then applied the model in a different cell line, H1-hESC, for prediction. Figure 2.5 shows the ROC curve for predicting CTCF, OCT4, and MAFK binding sites with the cross-cell-line-trained model. For MAFK (model trained on

IMR90) and OCT4 (model trained on mESC), Methylphet outperforms CENTIPEDE significantly. For CTCF on H1-hESC (model trained on IMR90 and mESC, respectively), Methylphet outperforms CENTIPEDE after the FPR is higher than 0.15, although CENTIPEDE performs slightly better before that. In terms of overall area under the curve, Methylphet is superior in all TFs. These results demonstrate that Methylphet achieves robust and precise prediction when the model trained in a different cell line, showcasing the broad utility of our method.

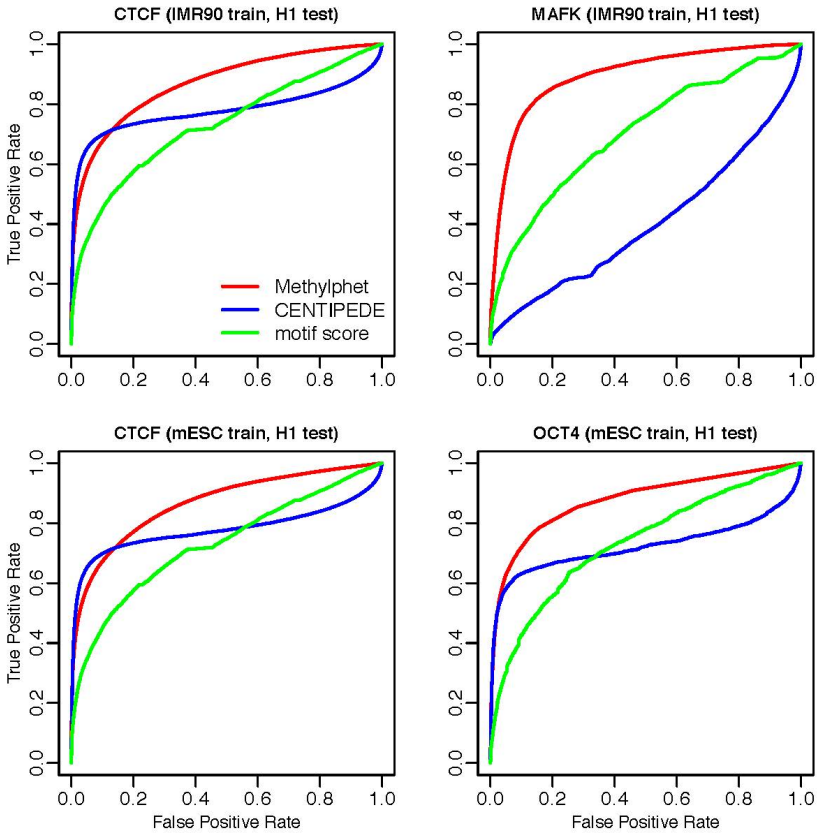

Figure 2.5 Cross-sample TF binding prediction results

Upper two figures are cross-cell line predictions between IMR90 and H1-hESC; Lower two figures are cross-cell line predictions between a mESC cell line and H1-hESC. The results demonstrate that Methylphet generally achieves more robust and precise prediction.

2.3.3 Cross-TF prediction results

We further investigated the TF-specificity of Methylphet model by cross-TF training and predicting . This result shows that even though cross-TF prediction is possible, TF-specific Methylphet model provides the best results. The methylation profile and other genomic characteristics of TF binding are important in the Methylphet model and better to be modeled in a TF-specific manner.

2.3.4 Experimental validation in mouse dentate gyrus (DG) cells

We performed NRSF binding site prediction in mouse dentate gyrus (DG) cells using Methylphet model trained from mES data. Because NRSF ChIP-seq data in mouse DG cells are not available, we performed qPCR in randomly selected sites as validation. Ten positive and ten negative sites are randomly selected from top 1000/bottom 1000 Methylphet-predicted binding sites respectively. Then five positive and five negative sites have suitable qPCR primers were tested. Fold enrichment is calculated on both positive sites and negative sites in order to compare the prediction performance. Among the selected sites, we can see clear enrichment inside positive predicted sites compared to negative predicted sites.

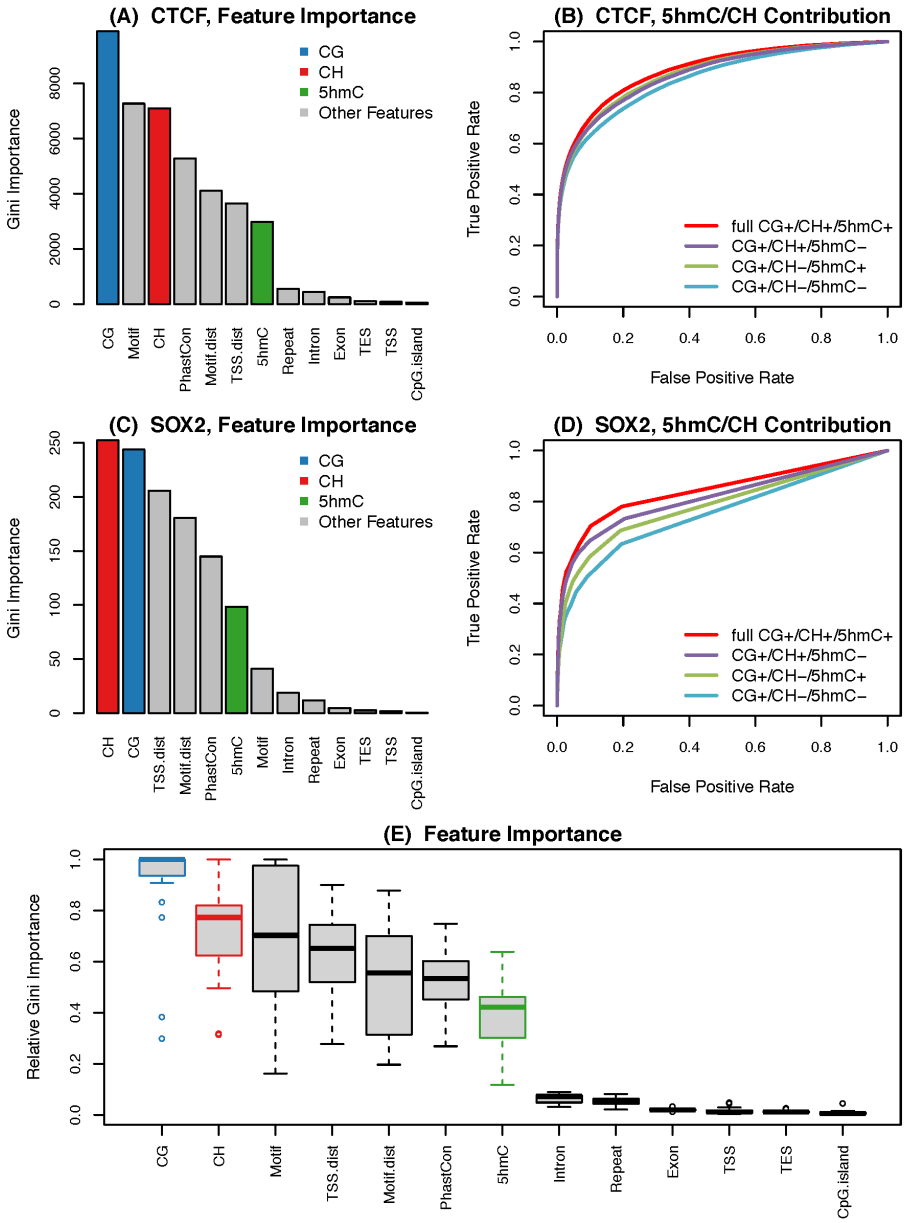

Figure 2.6 Relative predictive power of Methylphet features

(A) and (C) show the contribution of each feature in RF. (B) and (D) show the ROC curves with or without adding CG methylation and CH methylation information. Among them, (A) and (B) were generated using CTCF TFBS prediction results in H1-hESC; (C) and (D) were generated using SOX2 TFBS prediction results in H1-hESC. (E) Boxplot of Gini importance for each feature used in RF.

2.3.5 Contribution of different features in Methylphet

It is important to understand the relative predictive power of methylation levels and other genomic features used in Methylphet. We present the Gini importance of predictors in the RF model for CTCF and SOX2 using bar plots in Figure 2.6(A, C). Gini importance is the measurement of classification efficiency for each feature, which is defined as the level of decrease in the class impurity. Based on these analyses, we found that 5mC scores play the most important role in predicting CTCF binding, whereas the CH score is the most important predictor for SOX2. Motif is the second most important for CTCF, but its importance is very low for SOX2. This is because the SOX2 motif has much lower specificity compared to the CTCF motif. For both TFs, CH and 5hmC methylations play rather important roles. We compared the predictive performances with or without 5hmC and CH methylations. Figure 2.6(B, D) shows the ROC curves from such comparisons. These results demonstrate that including both 5hmC and CH methylation scores in the model improves the prediction power. Figure 2.6(E) shows the distributions of Gini importance of each predictor across all TFs we tested. We can see clearly that the 5mC score contributes the most on average among all features, next being the CH and motif scores, followed by sequence conservation and distance to the closest TSS.

2.3.6 Comparison with other predicting tools and other machine learning methods

We chose Random Forest as the ensemble algorithm to integrate all the features. During the RF model construction, one single variable was used at a time, and by integrating this information after sampling, it can give an automatic measure of feature importance. This

is important since our work requires integrating different types of information and evaluating feature importance. We also compared with other popular supervised machine learning tools, such as Neural Network (Venables, Ripley, & Venables, 2002), SVM (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel, 2010), and adaBoost (Culp, Johnson, & Michailidis, 2006). In general, we can obtain reasonably good results with all these choices because of the rich information in the methylation scores and other genomic features. Among all the methods we saw robust performance from RF across all the TFs. Even though the predicting result is not sensitive to model selection, we prefer RF for its additional advantages, such as its efficiency on large datasets, ability to avoid over-fitting, and its inherently non-parametric structure. In addition, it can provide more details in the importance of features without extra cost. The evaluation of Gini importance is done as the learning goes, which lead to one of the major discoveries in our study that 5mc and 5hmc profile can contribute as the top predictor in Methylphet.

2.3.7 Description of the software

R package Methylphet is freely available from https://github.com/stanleyxu/Methylphet and will be submitted to Bioconductor (Gentleman et al., 2004) soon. Methylphet accepts 5mC, 5hmc, and CH methylation profiles individually or in combination. As the example in the package shows, training about 7000 candidate sites and predicting on about 10,000 candidate sites with both CG and CH information takes less than one minute on a MacBook Pro laptop computer with 2.7 GHz i7 CPU and 16G RAM. Training time varies depending on number of candidate sites. For most of the cases in this study, training time is less than 30 minutes.

2.4 Discussion

In this work, we developed Methylphet, a novel computational method and software package to predict TF binding using a combination of methylation profiles and genomic features. The idea is based on the observation that *in vivo* TF binding events often co-occur with altered methylation levels. Methods for *in silico* prediction of TF binding using epigenetics data have been proposed before, mostly based on histone ChIP-seq or DNase-seq data. Our method exploits methylation data instead, which is much easier to collect experimentally. In this respect, our method provides a more practical means of *in silico* TF binding prediction and will be more useful in the clinical setting.

We show that Methylphet performs very well in the cross-sample and even cross-species predictions. These results imply that a predictive model trained under a certain biological context can be applied for prediction in different samples, which is important because it indicates that the model building procedure (which is the most time consuming) only needs to be performed once, and then the model can be applied elsewhere for the same TF. It is important to note that the predictive models are TF-specific, i.e., each TF will have its own model. This is because around the binding sites of different TFs, both the methylation patterns and the genomic features are different (Figure 2.2).

Disruption of epigenetic processes is known to contribute to the pathogenesis of multiple human diseases. For example, aberrant epigenetic modifications occurring at the earliest stages of neoplastic transformation are believed to be an essential player in cancer initiation and progression (Kanwal & Gupta, 2012; Verma & Srivastava, 2002). Using our method, a change of epigenetic status, particularly DNA methylation status, at a given locus could imply dynamics of *in vivo* TF-DNA interactions. Advances in epigenetics have not only

offered a deeper understanding of the mechanisms underlying disease pathogenesis, but have also allowed the identification of putative epigenetic biomarkers for early detection and diagnosis. Nevertheless, it would be very challenging to collect patient tissues/cells that are fresh enough to perform chromatin immunoprecipitation or DNase I-hypersensitive assays. However, DNA methylation analyses could be performed routinely with clinical samples. So the development of a DNA methylation-based *in vivo* TF-DNA interaction predicting algorithm is critical for uncovering effective biomarkers for human diseases(Schubeler, 2015).

One constraint of the application of the method is that whole-genome BS-seq experiment is very expensive. However, with the predictive model pre-built from public data, it is possible to use BS-seq data from selected regions (such as reduced representation of BS-seq, or RRBS (Meissner et al., 2008)) for binding prediction. Such an approach, although the prediction will not be genome-wide, still provides valuable information at important regions. Potentially, with small modifications of the methylation model, data from methylation microarrays can be used for binding prediction. This will be our research plan in the near future.

Unlike plant genomes, where enzymes for generating and erasing CH methylation have been well characterized(Heard & Martienssen, 2014), CH methylation in mammalian genomes has not been studied extensively until recently. Recent whole-genome bisulfite sequencing revealed that CH methylation is abundant in hESCs and hiPSCs, as well as brain(Lister et al., 2013). In brain, CH methylation accumulates during neuronal maturation, suggesting a potential role for CH methylation in normal brain function(Guo et al., 2014). The role of CH methylation in gene regulation remains elusive. Our analyses presented

here suggest that CH methylation is among the best predictors of *in vivo* TF-DNA interactions along with 5mC, pointing to an active role for CH methylation in gene regulation. It is possible that the coordination between CpG and CH methylations regulate the dynamics of TF-DNA interaction *in vivo*.

5-hydroxymethylcytosine (5hmC) shares similar characteristics with 5mC data. Since our model is very robust to capture the methylation pattern between TF binding sites and non-binding sites, we extended our model to summarize 5hmC data to calculate 5hmC score. Although 5hmC data are far more sparse than 5mC data, they could provide additional information to predict TF binding sites. Figure 2.6 shows the ROC curve with and without the 5hmC data.

Reference

Arvey, A., Agius, P., Noble, W. S., & Leslie, C. (2012). Sequence and chromatin determinants of cell-type--specific transcription factor binding. *Genome research, 22*(9), 1723--1734.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., . . . Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell, 129*(4), 823-837. doi:S0092-8674(07)00600-9 [pii]

10.1016/j.cell.2007.05.009

Bernat, J. A., Crawford, G. E., Ogurtsov, A. Y., Collins, F. S., Ginsburg, D., & Kondrashov, A. S. (2006). Distant conserved sequences flanking endothelial-

specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. *Human molecular genetics, 15*(13), 2098--2105.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Celniker, S. E., Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., . . . mod, E. C. (2009). Unlocking the secrets of the genome. *Nature, 459*(7249), 927-930. doi:10.1038/459927a

Consortium, E. P., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:10.1038/nature11247

Cooper, G. M., & Hausman, R. E. (2013). *The cell : a molecular approach* (6th ed.). Sunderland, MA: Sinauer Associates.

Cuellar-Partida, G., Buske, F. A., McLeay, R. C., Whitington, T., Noble, W. S., & Bailey, T. L. (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics, 28*(1), 56--62.

Culp, M., Johnson, K., & Michailidis, G. (2006). ada: an R package for stochastic boosting. *Journal of Statistical Software, 17*(2), 1-27.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data Via EM Algorithm. *Journal of the Royal Statistical Society Series B-Methodological, 39*(1), 1-38.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2010). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*.

Feng, H., Conneely, K. N., & Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res, 42*(8), e69. doi:10.1093/nar/gku154

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol, 5*(10), R80.

Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., . . . Song, H. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci, 17*(2), 215-222. doi:10.1038/nn.3607

He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., . . . others. (2010). Nucleosome dynamics define transcriptional enhancers. *Nature genetics, 42*(4), 343--347.

Heard, E., & Martienssen, R. A. (2014). Transgenerational epigenetic inheritance: myths and mechanisms. *Cell, 157*(1), 95-109. doi:10.1016/j.cell.2014.02.045

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., . . . others. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics, 39*(3), 311--318.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., . . . others. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods, 6*(4), 283--289.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., . . . Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res, 34*(Database issue), D590-598. doi:10.1093/nar/gkj144

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., . . . Zhu, H. (2013). DNA methylation presents distinct binding sites for human transcription factors. *Elife, 2*, e00726. doi:10.7554/eLife.00726

Ji, H., Li, X., Wang, Q.-f., & Ning, Y. (2013). Differential principal component analysis of ChIP-seq. *Proceedings of the National Academy of Sciences, 110*(17), 6789--6794.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet, 13*(7), 484-492. doi:10.1038/nrg3230

nrg3230 [pii]

Kanwal, R., & Gupta, S. (2012). Epigenetic modifications in cancer. *Clin Genet, 81*(4), 303-311. doi:10.1111/j.1399-0004.2011.01809.x

Klose, R. J., & Bird, A. P. (2006). Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci, 31*(2), 89-97. doi:S0968-0004(05)00352-X [pii]

10.1016/j.tibs.2005.12.008

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., . . . Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science, 341*(6146), 1237905. doi:10.1126/science.1237905

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell, 133*(3), 523-536. doi:S0092-8674(08)00448-0 [pii]

10.1016/j.cell.2008.03.029

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., . . . Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature, 462*(7271), 315-322. doi:nature08514 [pii]

10.1038/nature08514

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., . . . Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res, 42*(Database issue), D142-147. doi:10.1093/nar/gkt997

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., . . . Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature, 454*(7205), 766-770. doi:10.1038/nature07107

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research, 21*(3), 447--455.

Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., . . . Chinnaiyan, A. M. (2010). HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics, 11*, 369.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., . . . Ren, B. (2013). RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS computational biology, 9*(3), e1002968.

Ramsey, S. A., Knijnenburg, T. A., Kennedy, K. A., Zak, D. E., Gilchrist, M., Gold, E. S., . . . others. (2010). Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics, 26*(17), 2071--2075.

Rhee, H. S., & Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell, 147*(6), 1408-1419. doi:S0092-8674(11)01351-1 [pii]

10.1016/j.cell.2011.11.013

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., . . . others. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods, 4*(8), 651--657.

Schubeler, D. (2015). Function and information content of DNA methylation. *Nature, 517*(7534), 321-326. doi:10.1038/nature14192

Song, C. X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S. Q., Lin, L., . . . He, C. (2013). Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell, 153*(3), 678-691. doi:10.1016/j.cell.2013.04.001

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., . . . Schubeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature, 480*(7378), 490-495. doi:10.1038/nature10716

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics, 16*(1), 16--23.

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet, 9*(6), 465-476. doi:10.1038/nrg2341

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., . . . others. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology, 23*(1), 137--144.

Venables, W. N., Ripley, B. D., & Venables, W. N. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Verma, M., & Srivastava, S. (2002). Epigenetics in cancer: implications for early detection and prevention. *Lancet Oncol, 3*(12), 755-763.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., . . . Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res, 22*(9), 1798-1812. doi:22/9/1798 [pii]

10.1101/gr.139105.112

Won, K.-J., Ren, B., & Wang, W. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology, 11*(R7).

Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Zhang, L., Kim, A., . . . He, C. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell, 149*(6), 1368-1380. doi:10.1016/j.cell.2012.04.027

Chapter 3
Multi-layer Ensemble Learning Model Accurately Predict Transcript Factor Binding Sites Using DNase-seq and RNA-seq Data

3.1 Introduction

The goal of this challenge is to predict cell-type-specific Transcription factor binding sites(TFBS) for 32 TFs across 14 different types of cell lines. Cell-specific training datasets include chromatin accessibility (DNase-seq), gene expression levels and ChIP-seq data as gold standard. Non-cell-specific features can also be extracted from the given hg19 human genome DNA sequence.

TFBS can be affected by many factors. Sequence signature, or motif, is a major component for binding specificity. However, the ability to accurately define a TFBS varies significantly among different TFs. Also in different cell lines, motif will change for the same TF. In this challenge, I detect all potential motifs using de-novo motif discovery tool, MEME, with ChIP-seq data. Different motifs derived from various training cells are used together as DNA sequence features. DNase signal peaks are commonly considered as a union of major binding events. However the pattern of the DNase signals may vary among TFs. Gene expression level will indicate regulatory activities in the neighboring regions, thus in this challenge I will combine the gene expression levels together with the distance from a candidate region to the gene, in order to evaluate the potential role of TF regulation targeting at this gene. Random forest is used as the model to balance the contribution from heterogeneous features. For an

extremely unbalanced training set, ensemble methods with repeating sampling of negative sets were used to solve this issue. Each random forest model is trained within one cell. Final prediction is made by pooling predictions from all available training cells.

3.2 Methods

3.2.1. Feature Engineering of models

There are three parts of the feature space: Motif sets, DNase profile, and Expression of neighboring gens. The candidate regions for training and testing are 200 bp tiling bins across the majority part of human genome. There are 8,843,011 regions from Chr1, Chr8 and Chr21 as the hold-out test regions (referred as "ladder regions" below) in Ladderboard stage. There are 51,676,736 regions from the rest chromosomes as the training regions (refered as "train regions" below). The Final stage requires prediction on all these regions. I'll refer these regions as candidate regions in the following part of this write up.

For each region, a vector of length d is used as the feature vector to characterize this region. $d = 5n + 8 + 6$ in my current models, which n is the number of available training cell types. For example, for ATF2, there are 3 training cells: GM12878, H1 and MCF7. So the feature vector to characterize each ATF2 region will be of length $d = 5 \times 3 + 8 + 6 = 29$. Details for the composition of the features are described in the coming subsections.

It should be noted that the nature of these features are different: Motifs are non-cell-specific features, but DNase landscape and gene expression are associated with different cells. Thus when composing feature matrix for each TF-cell combination, the motif sets

matrix is the same for all the combinations with the same TF; but DNase matrix and Neighboring gene matrix should be changed in accordance with the cell.

3.2.1.1 Motif sets for each TF

Conservative peaks from ChIP-seq data were used to detect motifs within different cell lines. In each cell line, all peaks are ranked by the fold change of the signal, and top 1000 peaks were chosen for motif calling. MEME was used to call 5 significant motifs for each cell line. For a TF with n training cell lines, there will be $5n$ motifs used in the model. Called motifs were scanned across the whole genome using *matchPWM()* function from *Biostrings* package. Genome coordinates of hits, and their scores were saved. For all the train/ladder/test regions, each region will be assigned a vector of length $5n$ motif scores; if there are no motif hits, score is assigned 0; if there are multiple hits, score is the max of all hit scores. For a region set of length l, the final motif matrix will be of dimension $l \times 5n$.

For example, ATF2 has three training cell types: GM12878, H1 and MCF7. Using top 1000 conservative peaks from ChIP-seq data for ATF2 in GM12878, I ran MEME to call 5 motifs from each of these cell lines. The result will be 15 motifs, represented in 15 position probability matrices.

Next step, I scan the whole genome to find the matching sequence with each of these 15 motifs. For each motif (each PPM), I used *matchPWM()* function from *Biostrings* package to scan the whole genome. The result of this scanning process will be a list of target sequences, with their start/end genome coordinate, and a score (scaled 0~1) associated with the motif. The length of the matching target sequences are the same with

the length of the motif. This score describes how similar it is between the target sequence and the motif, and the minimum score for a matching process is 0.8. Usually there will be millions of hits across the genome (depending on the length of motif), but definitely not all across the whole genome.

The next step is to assign motif score to each regions. Naively I can perform motif scanning (the previous step) for all the regions, but that is not efficient (in fact, infeasible on my computing environment), since all the candidate regions are overlapping with each other. Alternatively, I only identify regions overlapping with motif matches, and assign motif score to these regions. For candidate regions with no overlaps with motif matches, the motif score is assigned 0. If there are multiple motif matches, the maximum score is selected.

The result from this stage is a motif set matrix. For example, for all the ladder regions (8,843,011 regions) for ATF2, the motif sets matrix is a matrix of dimension 8,843,011 * 29; Similarly, we can get motif sets matrix for train regions (dimension 51,676,736 * 29). Motifs are non-cell type specific. The information comes only from sequences and the called motifs. Thus, no matter which cell line are we going to use, motif sets matrix will be only calculated once for each TF.

3.2.1.2 DNase profile for each cell line

DNase signal intensity was summarized from bigwig files with *rtracklayer* package. For each train/ladder/test regions of 200bp, eight 25bp bins were opened within each region. For a region set of length l, the final DNase matrix will be of dimension $l \times 8$. The motivation for this 8-dim score vector, instead of a single DNase score, or peak/non-peak

binary classification is that the shape of DNase profile contains more information and the potential to capture TF-specific patterns.

DNase profile are cell-type specific, thus for each training cell, DNase matrix should be retrieved from the corresponding DNase profile.

For example, during the training of ATF2 model, using training cell line GM12878, we retrieve DNase profile from the bigwig file for GM12878 DNase-seq datasets. Take the first region of the ladder regions as an example, the genome coordinates for the first region is: chr1:600-800. The coordinates for the 8 bins for this region is: 600-624, 625-649, 650-674, … 775-799. DNase within each of these 8 bins were retrieved, and the result will be a vector of length 8, representing the DNase profile for this region. Finally, the DNase matrix of GM12878 for all the ladder regions will be a matrix of dimension 8,843,011 * 8. Similarly, DNase matrix of H1 or MCF7 can be constructed, but with the corresponding DNase datasets.

3.2.1.3 Gene expression levels and Distance to TSS

Linking the candidate regions to the target regulated gene is not a trivial issue. In this challenge I use the distance to the nearest three TSS together with expression level, in order to evaluate the potency of a region to be the binding site of a cis-/trans-regulatory factor.

TSS genome coordinates are extracted from the given GTF file. Distance is calculated in the 5'-3' direction: negative values indicates that the region is on the upstream side of the TSS (regardless strand). This distance comes from genome DNA, thus it is non-cell-

specific. Once the neighboring genes were found for each gene, and the distance is calculated, it can be saved for further use, and shared in different TF-cell combinations. For gene expression levels, TPM from two replicates are averaged as the measurement of expression. The reason for choosing TPM is to ensure that they are comparable among different samples. For a region set of length l, the RNA data matrix will be of dimension $l \times 6$, with 3 columns as the distances to the three nearest gene, and 3 columns as the expression level of the corresponding genes. Gene expression levels are cell-specific, thus in different training and predicting cells, gene expression levels should be retrieved accordingly.

3.2.2 Multiple Layer bagging random forest.

The final prediction result comes from 2 or 3 layers of bagging model, depending on the availability of multiple training cell types. At the first layer, all the input features (motif sets, DNase profile, TSS distances and corresponding gene expression level) are all fed into a random forest model. This procedure is repeated for 10 times, each time with a different sampled negative sets. In the second layer of bagging, 10 prediction results from these models were averaged. Finally, if there are multiple training cell types are available, predictions from different training cells were pooled together based on DNase similarity.

3.2.2.1 Training

The input of the training dataset is the combination of all the above three matrices. 500 trees with 1/3 sampled features were used to construct the forest. Imbalanced-class is a major problem in this challenge. Less than 1% of all the regions are true binding sites for

the majority of TFs. I used an ensemble learning strategy to train models within each training cell type. For each cell-TF combination in training, all the regions labeled "B" are treated as positive set. From the regions labeled "U", equal-sized regions were sampled, and one random forest model can be trained. This process is repeated for 10 times. So for each cell-TF combination, there will be 10 models trained, with differently sampled negative set. The motivation for this ensemble learning strategy is to deal with class-imbalance issue, while mutually cancel out noise due to sampling bias by averaging results from 10 models. For one TF with m training cell types, there will be $10m$ models trained.

For example, in the case of ATF2 model training, there are 3 types of training cell available: GM12878, H1-hESC and MCF-7. I trained 10 random forest models separately within each of these cell types. Within GM12878 cell, the input matrix is constructed by collecting: (1) ATF2 motif sets matrix, (2) DNase matrix from GM12878 DNase-seq data (3) RNA matrix from GM12878 RNA-seq data. The labels for ATF2 ChIP-seq in GM12878 is used as the gold standard. The training set contains balanced classes of binding ("B") and unbinding ("U") regions, with all regions labeled "B" as positive class, and sampled equal-sized "U" regions as negative class. After training, 10 GM12878-trained random forest models are ready for prediction. Similar steps can produce 10 H1-hESC models and 10 MCF-7 models.

3.2.2.2 Prediction

For each cell-TF combination to be predicted in the ladder stage, prediction was made using all the trained models from all training cell types. For one TF with m training cell

types, prediction results from 10 models within each training cell type is averaged; then the weighted sum of these m averaged result were used as the final prediction score. The weight is assigned based on the local similarity of DNase profile. This is based on the motivation that similar chromatin landscape might hint similar functionality during regulation. DNase profile similarity is calculated as the correlation coefficient between the tested cell type and the m training cell types.

For example, In the case of ATF2 binding prediction within K562. Firstly, input matrix is constructed by collecting: (1) ATF2 motif sets matrix, (2) DNase matrix from K562 DNase-seq data, (3) RNA matrix from K562 RNA-seq data. Then the input matrix is fed into all the 30 trained models, resulting in 30 prediction scores for every predicting region. Prediction scores within the same training cell are averaged, resulting in 3 scores: GM12878 score, H1-hESC score and MCF-7 score. Finally, these three scores are weighted average based on the DNase profile similarity between the training cells and K562. To be more specific: the more similar the local DNase landscape within this region between GM12878 and K562, the more weight the GM12878 score will get.

Prediction for the final round adopts the same strategy, except that applying on a larger test sets.

3.3 Discussion

The advantage of my method is the benefits from multiple layer's bagging. Since the complete region sets are too huge, it is almost certain that training should be based on sampled training sets. Across different sampling negative sets, averaging results from 10 models will cancel out error predictions due to sampling bias, while enhance convincing

prediction results. The class-imbalance issue can be solved by multiple sampled negative sets. Pooling results from different training cell lines based on chromatin similarity further weighs more on regions that have similar regulatory landscape.

The major concern for my current result is for TFs with weak motifs. Usually models trained for these TFs turn to rely more on DNase signals. This will directly lead to higher false positive rate in the prediction result, since DNase peaks are not TF specific. To further improve the performance of this method, higher resolution of DNase signal anchored at motif matching center will be a good direction. Alternatively, multi-task prediction might help to reduce false-positive rate, if we can tell one region is more likely to be bound by a different TF. Adding correlations between TF bindings is essential; However, the challenge only provide partial TF-cell combinations. Thus, more elegant modeling is required to characterize the comprehensive interaction map between TFs.

Chapter 4
Regulatory annotation of genomic intervals based on tissue-specific expression QTLs

4.1 Introduction

A large number of high throughput experiments have been producing results that can be summarized as a list of genomic intervals. For example, peaks from ChIP-seq (Barski et al., 2007; Johnson, Mortazavi, Myers, & Wold, 2007; Qin et al., 2010; Robertson et al., 2007), humps from ATAC-seq (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013; Buenrostro, Wu, Chang, & Greenleaf, 2015), DNase-seq (Song & Crawford, 2010), differential methylated bumps from WGBS (Jaffe et al., 2012; Lister et al., 2008; Wu et al., 2015), or linkage disequilibrium (LD)-spanned neighborhood around significant disease-associated single nucleotide polymorphisms (SNPs) identified from Genome Wide Association Studies (GWASs) (Welter et al., 2014). Typically, thousands or tens of thousands of such intervals are in the list, hence it is impossible to explore them one by one. How to effectively and efficiently discover biological properties and reveal biological insights from these large number of genomic intervals is an important yet challenging task. A common practice for interpreting such findings is a two-step process. First, link each of the genomic interval to its nearest gene, then study the properties of the list of genes derived from all the intervals, typically using methods such as gene ontology (GO) (Ashburner et al., 2000) term enrichment analysis (D. W. Huang, Sherman, & Lempicki, 2009) and gene set enrichment analysis (GSEA) (Subramanian et al., 2005). Examples of such approach are GREAT (McLean et al., 2010) and Enrichr (Kuleshov et al., 2016). The rationale

behind such a method is because most of the biological knowledge we have collected so far focus on genes. A drawback of the aforementioned approach is that many of the genomic intervals are tens or even hundreds of kb away from its nearest gene hence the assignment will be difficult to justify. Recent findings from chromosomal conformation capture-based technologies have showed that a distance regulatory element may be put in touch with its target gene by chromosomal looping (Smemo et al., 2014). Conversely, in gene-dense genomic regions, typically multiple genes can be found within the 10kb radius of a variant, making assigning the target gene by proximity comparable to random guessing. More importantly, since many genes exert their functions in a context-specific and tissue-specific manner, their property may not always be transferable to its nearby genomic interval. Both gene expression and epigenetic regulation are dynamic across different cell and tissue types (Torres et al., 2014; Wang et al., 2016), and such information is not considered using distance-based methods. Therefore, it is of great interest to explore the function of genomic intervals beyond inferring their functions simply from its closest gene. Rather than using nearby genes as surrogates, an alternative strategy is to explore the enrichment of functional elements inside these genomic intervals. Examples include DNA sequence motifs, DNA conservation and CpG Islands. More recently, thanks to large consortium efforts such as ENCODE (Bernstein et al., 2012), REMC (Bernstein et al., 2010) and IHEC (Stunnenberg, Consortium, Hirst, International Human Epigenome Consortium, & Hirst, 2016), increasing number of functional elements such as transcription factor *in vivo* binding sites detected by ChIP-seq experiments have been systematically cataloged. One can check whether their genomic intervals of interest are enriched with any type of functional elements (Griffon et al., 2015). The same idea can be extended to other types of

biological meaningful genomic entities. For example, checking the enrichment of trait-associated SNPs (taSNPs) identified by GWAS can help link the set of genomic intervals to diseases or traits (Chen & Qin, 2016).

Similar to GWAS-identified taSNPs, another important type of functional variants are the expression quantitative trait loci (eQTL), which are variants that shown association with the expression level of their target genes. Transcription regulation is one of the most important types of functional annotation in the non-coding part of the genome, and eQTL provide direct evidence of such connection. The eQTL catalog has seen significant boosted in recent years thanks to the Herculean effort of the Genotype-Tissue Expression (GTEx) consortium (Aguet et al., 2017). GTEx provides a comprehensive eQTL catalog with high quality and sufficient power given the large sample sizes and large number of tissue types. We felt that GTEx eQTLs provide a remarkably valuable resource to include in enrichment analysis for genomic intervals, because it provides a putative link to the genes these loci potentially regulate. Two aspects of the eQTL data can greatly improve the discovery of functional links between genomic loci and genes. Firstly, the association between loci and genes does not depend on genomic distance, which has been shown to be unreliable (see Result). Secondly, unlike proximity-based gene assignment which is static, eQTL information is tissue-specific, in that an eQTL may regulate its target gene(s)—referred to as eGenes from now on, only in one or two specific cell types. Which make the biological interpretation much more specific and informative.

We believe that studying the enrichment of eQTLs at the pathway or gene set level, instead of at the individual gene level, is necessary. In any given tissue, only about 20-30 eQTLs are found for each eGene in GTEx on average, and most of them are located near their

eGene—since GTEx primarily focused on *cis*-eQTL. Therefore, eQTLs for a single gene do not spread out of the neighborhood of their eGene therefore not suitable to test the enrichment of eQTLs for each gene individually. On the other hand, most pathways or gene sets contain tens to hundreds of genes that are functionally related. Hence, checking the enrichment of eQTLs whose eGenes belong to the same pathway or gene set can help us to potentially build connections between the genomic intervals and pathways or functional gene sets which will be highly informative. Some recent works (Ahmed et al., 2017; Li et al., 2016) show potential to perform functional enrichment analysis using eQTL data. However, a systematic evaluation of tissue-specificity is still lacking. In this work, we describe *loci2path*, a computational tool as an R Bioconductor package to enables straightforward enrichment analysis of eQTLs of pathways/gene sets for a set of genomic intervals. The current version of loci2path utilizes the entire eQTL catalog from the GTEx v6p data release which contains 1,702,612 unique eQTLs associated with 16,562 unique eGenes identified from 44 tissue types (Table S1). As of pathways and gene sets, the current version of loci2path contains 6,320 pathways from MSigDB (Liberzon et al., 2011) belonging to the BioCarta, KEGG and GO categories. To illustrate the utilities of loci2path, we test various trait-/disease-related genomic regions constructed from immune-related disease database immunoBase (www.immunobase.org) as query regions.

4.2 Result

4.2.1 Overview of *loci2path*

The components and workflow of loci2path are shown in Figure 4.1. We utilize eQTL data from the GTEx project, and pathway information from MSigDB for this study. In the

beginning, loci2path takes in a set of genomic regions as input. We next count the total number of eQTLs that fall in these intervals for each pathway (or gene set) and tissue type combination. This is then followed by enrichment test and tissue-specificity evaluation. In order to evaluate tissue specificity, we calculate the frequency that an enriched pathway is detected across all tissues. The Result contains two piece of information: (1) enriched pathways connected to the queried loci, ranked by enrichment; (2) the tissue type in which this enrichment is detected. User can customize the query result by sorting pathways based either on enrichment test p-values or degree of tissue-specificity (DTS, See Methods). Additional summary data, such as the numbers of eGenes or eQTLs, the size of the pathway, are also presented which may be used to filter the results.

In order to visualize both the pathway enrichment and the tissue specificity, loci2path presents the main query result as a heat map. The rows of the heat map are enriched pathways; the columns are tissues. Rows and columns are arranged by hierarchical clustering. The color for each cell indicates the degree of enrichment, rendered from red to blue as negative log p-values vary from high to low. We show in the following analysis that this visualization method helps to reveal interesting enrichment patterns, and offer clues on potential links between genetic variations and disease pathogenesis.

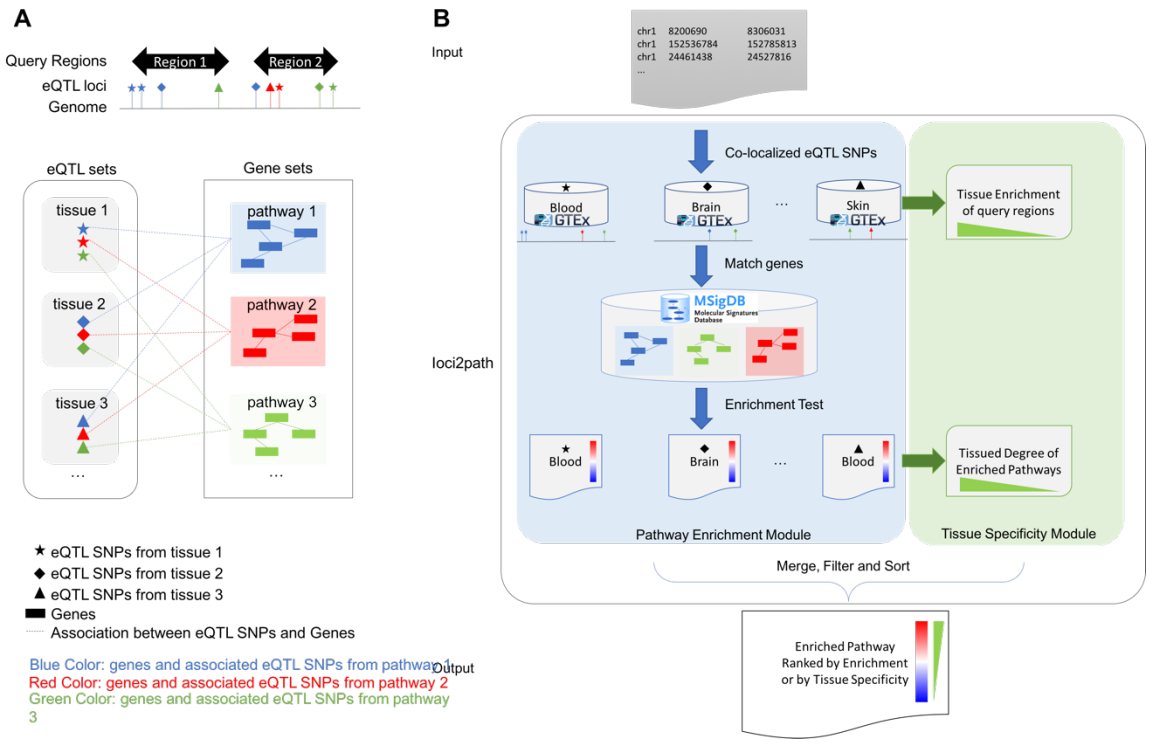

Figure 4.1 workflow of loci2path

(A)Component of *loci2path*. We use shapes to mark tissue or cell types, and colors to represent pathways. In the eQTL box, eQTL loci from different tissues were shown in shapes. In the Pathway box, genes from different pathways are shown in colors. Dash lines represents the association between eQTL and genes. (B)Workflow of *loci2path*.

4.2.2 GTEx eQTL data from 44 tissues

We downloaded all eQTLs from the GTEx data portal. The number of eQTLs and eGenes are summarized in Table S1. This dataset contains eQTLs identified from 7,051 samples representing 44 different tissue types collected from 449 donors (Aguet et al., 2017b). From Table 1 we noticed that the numbers of eQTLs and eGenes vary among tissues. The number of eQTLs ranges from 34,898 (Uterus) to 577,857 (Thyroid) and the number of associated

eGenes ranges from 542 (Vagina) to 6,990 (Tibia Nerv). The sample size is a major factor of the wide range of eQTL/eGene numbers (Table S1), though other factors such as tissue-specific gene expression (Aguet et al., 2017b) and post-mortem interval (Ferreira et al., 2018) might also contribute to such differences. The 44 tissue types contain clusters of homogeneous tissues, such as multiple types of brain, skin, muscle and artery cells.

Examining the relationships among eQTLs, eGenes and tissue types composition in GTEx data reveals that only about 25% of the times that the closest gene of an eQTL turned out to be its eGene (Figure 4.2A). This result highlights the danger of simply assigning genes to a locus based on genomic proximity. On the other hand, around 10% of the eQTLs are located in loci where multiple genes are located within a 10kb neighborhood, in which case assigning the target gene by proximity is very unreliable. Next, we explore the tissue-specificity of eQTLs and eGenes. The DTS for a gene is decided by the number of tissues in which this gene is detected as eGene (see Method). We find that around 30% of eQTLs are detected in only one tissue (i.e., tissue specific), while more than half of all eQTLs are detected in one or two tissues (Figure 4.2B). However, the proportion of tissue specific eQTLs vary drastically among different tissues. (Figure 4.2C).

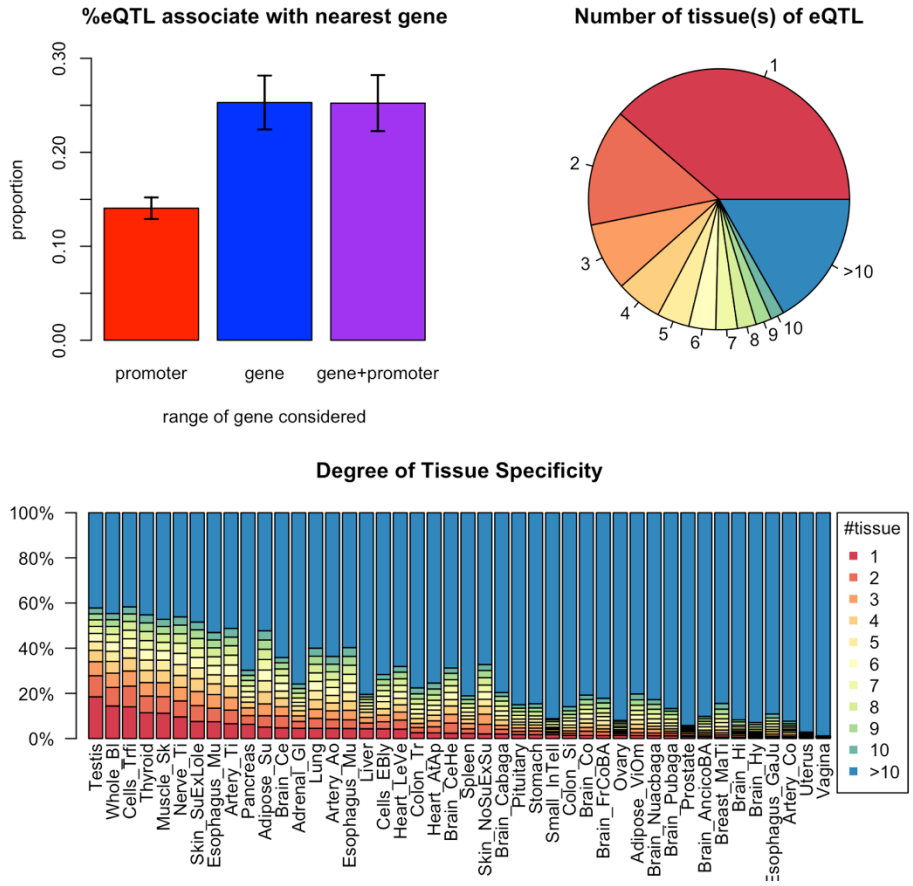

Figure 4.2 Tissue specificity of eQTL data

(A) percentage of eQTL associated with its nearest gene. Three bars represent different ways to define nearest gene. For one gene-eQTL pair, three types of distance are considered: (1) distance to gene promoter (defined as -2000 ~ +200bps of TSS); (2) distance to gene body (TSS~TES); (3) distance to promoter + gene body (-2000 of TSS ~ TES). (B) Number of tissue(s) each GTEx eQTL is associated with, ranging from 1(only detected as eQTL from one tissue) to 44, total number of tissues from GTEx collection. (C) Distribution of DTS of eQTLs within each tissue. Inside the bar plot, each bar shows the composition of eQTLs with different DTS. Tissues are ordered with an increasing

average DTS. On the left are the ones having more tissue-specific eQTLs; on the right are tissues having non-tissue-specific eQTLs.

4.2.3 MSigDB Pathways

We collected pre-defined gene pathways from the MSigDB (Liberzon et al., 2011; Subramanian et al., 2005). In this study, we query loci2path for enrichment of three categories of MSigDB pathways: GO terms, KEGG pathways and BioCarta pathways. GO terms are commonly known as the most comprehensive resource for functional annotation of genes. In this study, we collected all three GO gene sets from MSigDB's class c5 gene sets including BP-biological process, 4436 sets; CC- cellular component, 580 sets; MF-molecular function, 901 sets. In addition, we included BioCarta pathways to accommodate more details of interactions among gene members regardless the hierarchical relationships of gene sets that are comprehensively defined in GO. In total, 217 BioCarta gene sets from MSigDB's c2: curated gene sets were downloaded. We also collected 186 KEGG pathways to detect metabolism related functions. Details of the pathway resources are listed in Method section.

4.2.4 Query regions from immunoBase

We first analyze immune-related diseases. For each disease. We use risk regions defined by immunoBase (https://www.immunobase.org/) as the input. ImmunoBase provides a curated data source for immunologically related human diseases. This collection of findings from GWASs and fine mapping studies using the immunoChip serves as a valuable resource to study immunological disorders (Cortes & Brown, 2011;

Polychronakos, 2011). Then the neighborhood around each GWAS variant based on local linkage disequilibrium (LD) is added and overlapping neighborhoods are merged to form disease risk regions. For example, the input regions for Psoriasis are constructed by merging regions within the ±0.1centimorgan genetic linkage ranges around disease-related variants (https://www.immunobase.org/region/table/PSO/). Using these data, Chun et al. identified autoimmune-disease related risk signal enriched in gene regulatory regions in a tissue-specific manner (Chun et al., 2017). In our study, we conduct a pathway enrichment analysis leveraging eQTL information which allows us to study the tissue-specificity of functional enrichment within pathways. We choose to investigate all the 12 core immune diseases originally targeted by the immunoChip consortium.

4.2.5 Tissue specificity captures distinct modules of pathogenesis in Psoriasis

Psoriasis (OMIM ID: 177900) is a common, chronic skin disorder with a complex genetic and environmental etiology characterized by epidermal hyper-proliferation, vascular remodeling and inflammation (Nestle, Kaplan, & Barker, 2009). Many genetic studies (Greb et al., 2016; Gudjonsson & Elder, 2007; Hwang, Nijsten, & Elder, 2017) including GWAS studies (Strange et al., 2010) and meta-analyses (Tsoi et al., 2015) have been conducted and tens of genomic loci have been identified as Psoriasis-associated. In ImmunoBase, 45 loci are included in 35 regions covering a total of 10.69 million base pairs (MB). A heat map of highly enriched pathways with a filtering p-value threshold of 1e-4 for the Fisher's exact tests is shown in Figure 4.3.

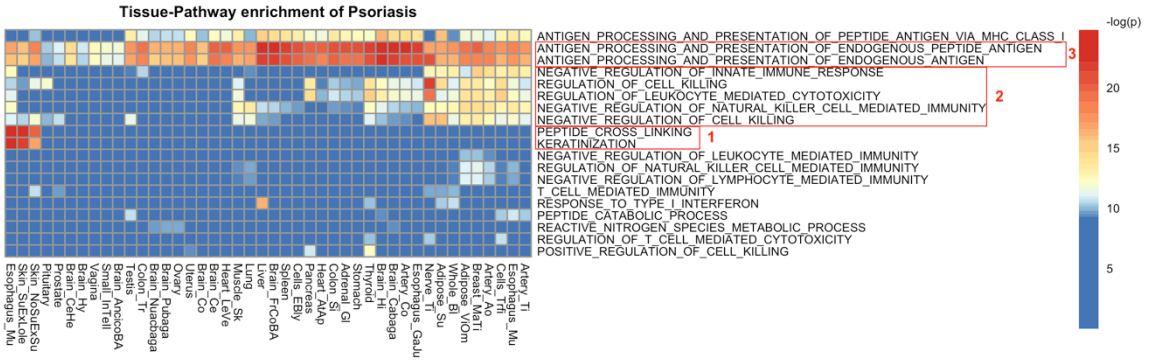

Figure 4.3 Heat map of Tissue-Pathway enrichment of Psoriasis

Rows of the heat map are pathways; columns are tissues. Each cell shows the significance of enrichment by –log(p-value). Red color indicates high enrichment, while blue indicates no enrichment. Three groups of pathways with distinct DTS are highlighted with red boxes and numbered as group 1~3.

From the heat map shown in Figure 4.3, we notice that three different categories of GO pathways showed significant difference in their enrichment patterns across tissue types. Pathway group 1 are enriched only in epidermis tissue types, including two skin tissues and mucosa. Pathway group 2 show enrichments in several tissue types that harbor dendritic cells with fuzzy block edge, but are absent from the majority of brain tissues. It is interesting to see that the majority of function is down regulating immune-response; and dysfunction of such pathways will cause autoimmune diseases such as Psoriasis. Pathway group 3 are all MHC I peptide presentation pathways, and completely non-tissue specific. The pathways of each group are listed in Table 4.1.

Table 4.1 Enriched pathway groups of psoriasis risk regions

Group	Pathways	Shared eGenes
1	Keratinization Peptide cross linking	LCE3C, LCE3D, LCE3E LCE1E, LCE3A
2	Negative regulation of cell killing Regulation of cell killing Negative regulation of innate immune response Negative regulation of leukocyte mediated immunity Negative regulation of natural killer cell mediated immunity	HLA-B, MICA, LGALS9
3	Antigen processing and presentation of endogenous peptide antigen Antigen processing and presentation of endogenous antigen	HLA-B, ERAP2, HLA-C, ERAP1

In accordance with the patterns shown in the heat map, DTS analysis yield the same three categories (Figure 4.4B). We extract the genes within each category and discover distinct composition of gene members, and we also notice distinct clusters of gene functions within each group. In Figure 4.4C, the most frequent genes from each pathway group are shown. Group 1 are dominated by late cornified envelope family genes. Figure 4.4A use LCE cluster 3 genes as an example to show the spatial relationships among query region, eQTLs, eGenes and additional GWAS evidence on the genome. This shows how *loci2path* detect

the connection and perform the enrichment analysis. Multiple late cornified envelop genes locate in this region, and was defined as the cluster 3 genes. The deletion of the gene LCE3B and LCE3C has been identified as a risk factor for Psoriasis (De Cid et al., 2009) . However, within this gene-dense region, assigning target gene of the association based on distance will include irrelevant genes within the same genomic region. Additional GWAS studies identifying proximal loci associated with Psoriasis are also shown on the figure (see Methods). Group 2 includes genes involved in innate immune response. For example, MICA is the gene coding NK cell attracting peptide (Menier, Riteau, Carosella, & Rouas-Freiss, 2002); NOS2 encodes the cytokine inducible enzyme (Stuart et al., 2010); And LGALS9, a versatile factor in immune homeostasis, is reported to be down-regulated and dysregulate helper T-cell signaling in Psoriasis patient (De La Fuente et al., 2012; Golden-Mason & Rosen, 2017). Group 3 has some distinct members involved in antigen processing and presentation, such as ERAP1 and ERAP2, together with class 1 MHC encoding gene HLA-B and HLA-C. The non-tissue-specific feature is not surprising, since peptide presentation is a global event across all cells. Variations of these genes will result in the altered antigenic MHC complex that triggered the downstream T-cell activation associated with autoimmunity (Goris & Liston, 2012).

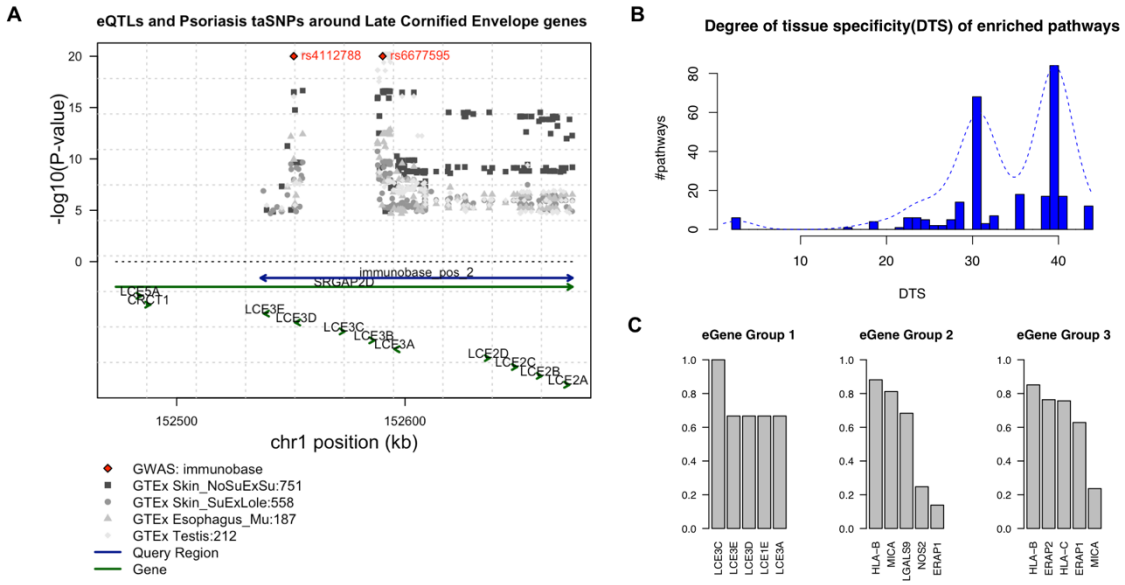

Figure 4.4 Degree of tissue specificity from Psoriasis

(A) LCE cluster 3 genes, an example to show the spatial relationships between query region, eQTLs, eGenes on the genome. Green arrows at the bottom are genes, with arrows showing the direction of genes. Blue double-arrow line at the bottom are the input query region. Diamond dots with colors are GWAS loci associated with Psoriasis, downloaded from immunoBase. Grey dots are GTEx eQTLs, with height as p-value in negative log scale. Different shapes and shades represent different tissue origin. (B) Distribution of DTS for enriched pathways using Psoriasis risk regions as query. We observed three clusters of DTS, which are in concordance with the clusters in the Tissue-Pathway heat map.(C) Most frequent eGenes from the three groups of enriched pathways using Psoriasis risk regions as query. Top five most frequent genes from each pathway group are shown. Name of the gene and its proportion appearing in the pathway group are shown.

After a systematic review of Psoriasis pathogenesis, we noticed these three categories are in concordance with the three major modules of the tentative model of Psoriasis, as described by Bergboer et. al (Bergboer, Zeeuwen, & Schalkwijk, 2012). Group 1 pathways relate to skin barrier and keratization module. Group 2 pathways relate to innate immune systems. Group 3 pathways are general immune response pathways of adaptive immune system. It is interesting to note that a well-studied epistasis between HLA and ERAP1 is captured in group 3 pathways (Bergboer et al., 2012; Goris & Liston, 2012). HLA encode individual specific MHC, and ERAP1 code the enzyme involved in trimming HLA class I-binding peptide. Variation on this would affect whether the peptide can be presented to MHC1, thus revealing the mechanism of Psoriasis risk within certain population of a specific HLA subtype.

We compared to the query result from GREAT (McLean et al., 2010), using the same set of query regions, and GO terms as pathway gene sets. We discovered that more than 50% of the top enriched pathways (with p-value < 1e-5) from GREAT are also discovered by loci2path. However, loci2path adds tissue specificity information from using eQTL data, which demonstrates the potential to combine the function of associated genes with their context of tissue type in the study of disease-related genomic regions. We noticed that the skin tissue specific pathways are not detected from GREAT query result.

4.2.6 Shared risk pathways among 12 core Immune Disease

Next we extend our query to all 12 core immune diseases from immunoBase. We again organize and present the results in heat maps with rows represent pathways, and columns represent the 12 immune diseases, in order to examine the inter-relationships among

complex immune diseases. We generate three such heat maps from three immune-related tissues: Blood, thyroid and Spleen. And we query against two collections of gene sets: BioCarta pathways and GO terms from MSigDB (see Methods).

From the heat maps, we observe interesting patterns. Firstly, enrichment patterns across 12 diseases show significant differences across the three tissue types, suggesting that tissue information is very important in eQTL studies. Among the three tissues, heat map from the blood shows the most enriched pathways, probably due to the rich repository of immune cells (leukocytes lymphocytes), in spite of a relatively smaller repository of eQTLs in blood compared to thyroid. One surprising example is for autoimmune thyroid disease; we are not able to find enriched pathways in Thyroid. Additionally, we found Ankylosing spondylitis and autoimmune thyroid disease show distinct patterns from the other ten diseases. Further examination of the query regions for these two diseases shows that these regions does not include the HLA gene complex, which resides on a 3Mbp stretch within chromosome 6p21, while other immune diseases have risk regions overlapping the HLA complex region. Thus the heat map results for these two diseases show distinct patterns of lacking immune-related pathways that other diseases share.

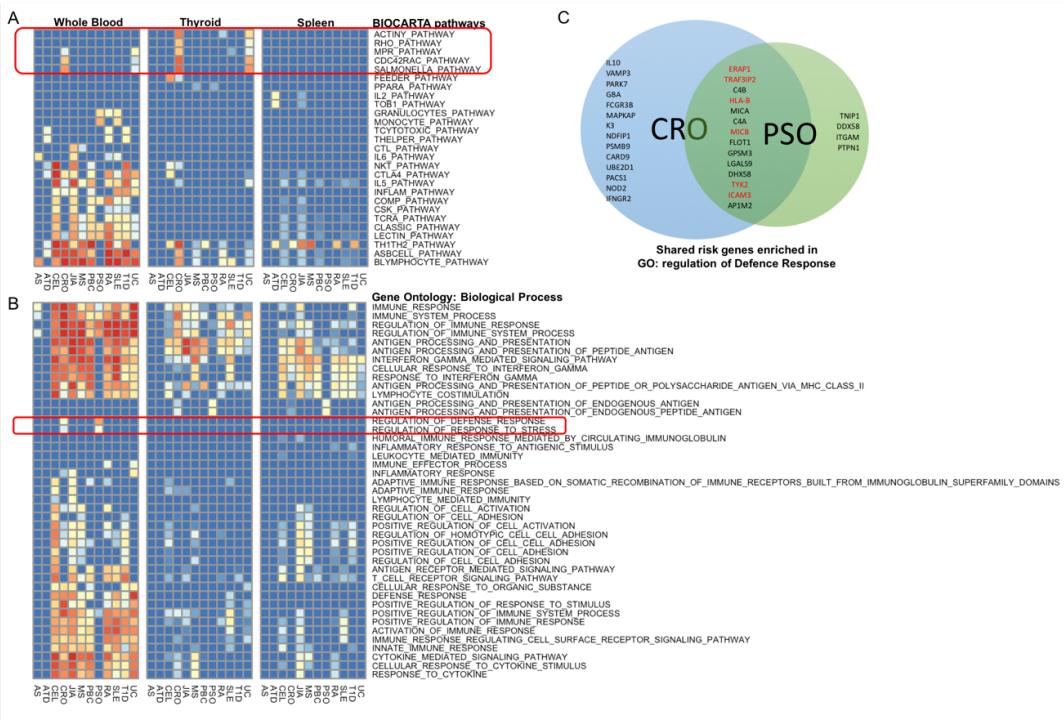

Figure 4.5 Heat map of 12 immune diseases

(A) Disease-Pathway heat map of enriched BioCarta pathways from three tissues. All the 12 immune diseases risk regions are queried against BioCarta pathway collection repeatedly in blood, thyroid and spleen, resulting three heat maps. Within each heat map, the rows are enriched pathways, and columns are diseases. IBD specific pathways are marked with the red box. (B) Disease-Pathway heat map of enriched GO pathways. Queries were performed in the same way as in Figure 5A. Two enriched GO Biological Process terms are Psoriasis specific and distinctly shared with CRO, and they are highlighted with a red box. (C) Venn diagram of gene members of the two distinctly enriched pathways. 27 genes form CRO and 18 genes from PSO are associated with eQTLs covered by disease risk regions. Among them, 14 genes are in common. Among the shared genes, the ones with literature report as the evidence of the shared risk genes are highlighted in red.

Within BioCarta pathways (Figure 4.5A), we observed similar patterns between two diseases: Crohn's disease (CRO) and Ulcerative colitis (UC). This is expected since these are two subtypes of inflammatory bowel disease (IBD) thus sharing numerous enriched pathways. Among the shared enriched pathways, Salmonella pathway is related to bacteria entering membrane of salmonella infected cells, which plays an important role in the onset of IBD (Henderson & Stevens, 2012; Schultz et al., 2017). In addition, several pathways associated with Rho family genes: RhoA, Rac and Cdc42 are commonly enriched as well. It has been reported that Rho kinase signal pathway is involved in the three-essential starting stage in the chronic pathogenic procedure of IBD (Y. Huang, Xiao, & Jiang, 2015). Also, pathway Actin Filament Y, one component of cytoskeleton, plays vital role in the disruption of epithelial barriers under inflammatory conditions (Ivanov, Parkos, & Nusrat, 2010).

In the enrichment of GO terms (Figure 4.5B), in addition to the similar patterns we described in BioCarta, we find two unique pathways that are shared only by CRO and PSO: defense response and response to stress. We map them onto the hierarchy of GO terms, and found defense response is a sub-ontology of response to stress. We extracted the member genes of the leaf node from these two diseases, and show them in the Venn diagram (Figure 4.5C). We found that 50% of CRO genes and 70% PSO genes are in common in this pathway. Further literature research reveals that there are several genes among them have been identified as common risk genes between CRO and PSO. For example, variants located in the TRAF3IP2 gene contribute to the susceptibility to immune diseases involving the skin or the gut (Ciccacci et al., 2013). JAK family kinase TYK2 functions as

mediator of IL-12 and IL-23, which are key factors in the pathogenesis of PSO and IBD. Effective inhibitor of TYK2 provides an attractive therapeutic strategy for such diseases (Miao, Masse, Greenwood, Kapeller, & Westlin, 2016). ICAM-3 plays an important role in neutrophils to amplify NK cells by producing Inteferon-gamma, with supporting evidence from both PSO and CRO patient samples (Costantini et al., 2011). The tissue location of neutrophils also suggests the reason that the enrichment within blood is the most significant among the three tissue types in the heat map. For the rest of MHC-peptide antigen presenting genes, these shared genes function in defense response pathways, and variants associated with these genes will affect the normal immune defense mechanism by altered antigenicity and immune regulatory pathways.

4.2.7 Software availability

R package *loci2path* is freely available from *github.com/StanleyXu/loci2path* and Bioconductor with package name '*loci2path*'. User can provide arbitrary query regions in R using the *GenomicRanges* data type in R. For query regions of 600kbps, using the complete 44 tissues GTEx eQTL set, query against the complete BioCarta pathway collection (217 gene sets) takes less than 1.5 minute to finish on a MacBook Pro laptop computer with 2.9 GHz i5 CPU and 8G RAM. Parallel query mode would further increase the speed on a multi-core computing platform, on which the performance varies due to the working load and availability of resource.

4.3 Discussion

We developed *loci2path*, a novel computational tool to annotate non-coding genomic regions using comprehensive tissue-specific eQTL information. Functional annotation of non-coding genomic regions relies on accurately mapping interested loci to other functional elements on the genome. The rationale of loci2path is to build the mapping via existing knowledge of tissue-specific association between genomic loci and genes. Compared with analysis based on single gene that harbor risk of false associations, enrichment analysis provide robust assessment of function by integrating multiple genes with pre-defined pathways and brings insight to biologically meaningful results. We believe that loci2path would help researchers to identify accurate functionality annotation and specific tissue enrichment for query regions of interest.

We perform enrichment analysis using eQTL data to link genes to genomic loci. In this study, eQTL data from GTEx and pathways from MSigDB were collected to study functional enrichment of pathways/gene sets for risk regions harboring variants associated with immune diseases. We discovered that DTS from *loci2path* query result reveals three different but corroborating underlying pathogenesis modules in the query of Psoriasis risk regions. We also discovers that pathways that show distinct enrichment patterns in CRO and UC compared to other immune diseases are involved in different ways of pathogenesis of IBD. In addition, we identified common disease risk factors from shared enriched pathways among the three tissue types: population HLA type, variation in antigen-presentation, and variation in innate immune response. This pattern shows more significant enrichment in blood, rather than the other two immune-related tissues, perhaps due to the large proportion of leukocytes participating in the immune-related diseases.

The software is written in R and published in Bioconductor. Vectored statistical tests and parallel processing makes it run ultra-fast to perform enrichment tests across eQTL data from multiple tissues. A standard S4 class data structure enables users to customize annotation resources. For example, one might extend the loci-gene connection from eQTL data to any arbitrary mapping relationships, such as known regulatory genomic loci and target gene. Similarly, users can define arbitrary genomic regions of interest as query input. The potential applications of loci2path for arbitrary query regions include, but not limited to: regions containing trait-associated SNPs identified by GWAS, regions showing differential methylation levels between two groups, regions harbor different groups of transcription factor occupancy that plays different regulatory roles on target genes in certain diseases. *loci2path* is designed to build links between any genomic region(s) and targeted gene(s), which is not based on distance.

We believe the accumulating eQTL data become increasingly useful as a rich information resource. Thanks to the accumulating diverse tissue and cell types and enhanced statistical power due to increasing sample sizes, the growing eQTL resource would greatly improve query quality of *loci2path*. Availability and user-friendly data portals are making research to explore these public resources with *loci2path* more and more convenient. Together with more refined pathways, there are more enrichment patterns for traits, diseases and health to be uncovered, and *loci2path* is a powerful tool in this task.

In the future, we plan to continue adding latest eQTL information to loci2pth. It is expected that consortia like GTEx will add more tissue types with increased sample size. Another feature could be implemented is the order information of the query regions. Users might

have different importance measurement assigned to each query region. However, it is difficult to directly transfer the order information from query regions to eGenes, because the order might be further affected by other factors, such as the order of eQTLs (ordered by association test statistics). Thus, enrichment methods for ranked genes can't be directly borrowed into our framework, and more elaborate modeling is required to reconcile order information from multiple resources.

4.4 Method

4.4.1 Enrichment measurement

For one eQTL set $\boldsymbol{ES_k}$ and one gene set $\boldsymbol{GS_j}$, we use the p-value from enrichment test (Fisher's exact test) to evaluate the significance of enrichment. The default enrichment test is carried out with gene based mode, in which *loci2path* will firstly identify the genes $\boldsymbol{g}$ associated with eQTLs from $\boldsymbol{ES_k}$ covered by the query regions, then evaluate the significance of enrichment of these genes $\boldsymbol{g}$ within a given gene set $\boldsymbol{GS_j}$.

4.4.2 Assessment of tissue specificity

Once the eQTL set list is ready, the tissue specificity for one eGene is evaluated by the number of tissues one gene is detected as eGene. For example, for one gene g_i, the degree of tissue specificity (DTS) is defined as:

$DTS(g_i) = \sum_k I(\, g_i \in \boldsymbol{ES_k})$,

where I is the identity function, $I(g_i \in \boldsymbol{ES_k}) = 1$ if g_i is eGene in tissue k, and $= 0$ otherwise.

4.4.3 Tissue Specificity measured by average tissue number

Due to the fact that the expression profile of different cell type result in different SNP-gene association, we are interested to know if the genomic region-pathway/gene set link (through eQTL) is specifically significant in one tissue only, or it is a global enrichment across multiple different tissue types. Previously we calculated tissue specificity score as DTS for each eGene already. In the query result, we define the average DTS as the measurement of tissue specificity for a given set of query regions and a specific gene set $\boldsymbol{GS_j}$.

$$avg\_DTS_j = \frac{1}{l_G} \sum_{g_i \in GS_j} DTS(g_i)$$

where l_G are the total number of eGenes from $\boldsymbol{g}$ that are members of gene set $\boldsymbol{GS_j}$.

4.4.4 Output

All the enrichment scores, counts used in the calculation and tissue/gene set identifiers are organized in a table as output. Each row of this result table contains data of a pair of $\boldsymbol{ES_k}$ and $\boldsymbol{GS_j}$. All the rows are ranked by p-values calculated from fisher exact test by default.

4.4.5 Tissue Enrichment test of query regions

The enrichment within tissues are tested before pathway enrichment. A binomial test was performed to determine the enrichment of the query regions within the tissue from which the eQTL loci were identified.

The probability of covering a eQTL loci in tissue k is noted as $p_k = l_k/l_{total}$, where l_k is the number of regions overlapping with eQTLs in tissue k, and l_{total} is the total length of the genome.

Given n query regions, the number of regions overlapping with eQTL loci were calculated as n_k. Binomial test calculates the p value of the enrichment in tissue k as the probability of having at least n_k regions overlapping with eQTL loci from tissue k:

$$\sum_{i=n_k}^{n} \binom{n}{i} p_k^i (1-p_k)^{n-i}$$

4.4.6 Multiple-test correction using adjusted p-value

Adjusted p-value is calculated using "BH" method(Benjamini & Hochberg, 1995) from function p.adjust() in R.

4.4.7 Datasets

GTEx eQTL

In this study, we collected eQTL sets from GTEx project, which are composed of eQTL studies from 44 type of tissues. GTeX eQTLs from 44 tissues were downloade from GTeX with command *wget* via the link: http://www.gtexportal.org/static/datasets/gtex_analysis_v6p/single_tissue_eqtl_data/GTEx_Analysis_v6p_all-associations.tar. Entrez ID is used as the default gene identifier. If the gene identifier is different between eQTL study and the gene set, they are all converted to Entrez ID. Unmapped genes are not included.

MSigDB pathways

Pathway and gene sets for this study were downloaded from MsigDB website: http://software.broadinstitute.org/gsea/downloads.jsp

Entrez ID was used as the identifier of genes across all the gene sets. This is based on the official release notes from MSigDB v3.1 (http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/MSigDB_v3.1_Release_Notes)

"Human Entrez Gene IDs and the corresponding symbols for the MSigDB v3.1 gene sets are based on gene_info.gz and gene_history.gz, downloaded from the Entrez Gene FTP site on November, 15, 2011"

Gene annotation

Genome Coordinates of all the entrez genes on reference genome hg19 were obtained using UCSC Known gene table, retrived with Bioconductor/GenomicFeatures package. This is the most updated version (downloaded Oct.24, 2016). There are in total 23056 genes, only 21063 of the MSigDB Entrez ID can be matched onto this set (~65%). By manually checking the missing genes, a majority of these records were withdrawn pseudo genes, thus excluded in the downstream analysis.

Gene expression

Tissue-specific gene expression level was obtained from GTEx data portal. The median RPKM table (GTEx_Analysis_v6p_RNA-seq_RNA-SeQCv1.1.8_gene_median_rpkm.gct) is used to quantify gene expression. Gene expression data for 44 tissues were downloaded from GTeX portal. File is named as "GTEx_Analysis_v6p_RNA-seq_RNA-SeQCv1.1.8_gene_median_rpkm.gct.gz". This file contains gene median RPKM by tissue.

Reference

Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., … Montgomery, S. B. (2017a). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. http://doi.org/10.1038/nature24277

Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., … Montgomery, S. B. (2017b). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. http://doi.org/10.1038/nature24277

Ahmed, M., Sallari, R. C., Guo, H., Moore, J. H., He, H. H., & Lupien, M. (2017). Variant Set Enrichment: An R package to identify disease-associated functional genomic regions. *BioData Mining*, *10*(1). http://doi.org/10.1186/s13040-017-0129-5

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*. http://doi.org/10.1038/75556

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., … Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. http://doi.org/10.1016/j.cell.2007.05.009

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*. http://doi.org/10.2307/2346101

Bergboer, J. G. M., Zeeuwen, P. L. J. M., & Schalkwijk, J. (2012). Genetics of psoriasis: evidence for epistatic interaction between skin barrier abnormalities and immune

deviation. *The Journal of Investigative Dermatology*, *132*(10), 2320–31. http://doi.org/10.1038/jid.2012.167

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74. http://doi.org/10.1038/nature11247

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., … Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), 1045–1048. http://doi.org/10.1038/nbt1010-1045

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. http://doi.org/10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, *2015*, 21.29.1-21.29.9. http://doi.org/10.1002/0471142727.mb2129s109

Chen, L., & Qin, Z. S. (2016). TraseR: An R package for performing trait-associated SNP enrichment analysis in genomic intervals. *Bioinformatics*, *32*(8), 1214–1216. http://doi.org/10.1093/bioinformatics/btv741

Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., … Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics*, *49*(4), 600–605. http://doi.org/10.1038/ng.3795

Ciccacci, C., Biancone, L., Di Fusco, D., Ranieri, M., Condino, G., Giardina, E., … Borgiani, P. (2013). TRAF3IP2 gene is associated with cutaneous extraintestinal manifestations in Inflammatory Bowel Disease. *Journal of Crohn's and Colitis*, *7*(1), 44–52. http://doi.org/10.1016/j.crohns.2012.02.020

Cortes, A., & Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Research and Therapy*. http://doi.org/10.1186/ar3204

Costantini, C., Calzetti, F., Perbellini, O., Micheletti, A., Scarponi, C., Lonardi, S., … Cassatella, M. A. (2011). Human neutrophils interact with both 6-sulfo LacNAc+ DC and NK cells to amplify NK-derived IFNγ: Role of CD18, ICAM-1, and ICAM-3. *Blood*, *117*(5), 1677–1686. http://doi.org/10.1182/blood-2010-06-287243

De Cid, R., Riveira-Munoz, E., Zeeuwen, P. L. J. M., Robarge, J., Liao, W., Dannhauser, E. N., … Estivill, X. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature Genetics*, *41*(2), 211–215. http://doi.org/10.1038/ng.313

De La Fuente, H., Perez-Gala, S., Bonay, P., Cruz-Adalia, A., Cibrian, D., Sanchez-Cuellar, S., … Sanchez-Madrid, F. (2012). Psoriasis in humans is associated with down-regulation of galectins in dendritic cells. *Journal of Pathology*, *228*(2), 193–203. http://doi.org/10.1002/path.3996

Ferreira, P. G., Muñoz-Aguirre, M., Reverter, F., Sá Godinho, C. P., Sousa, A., Amadoz, A., … Guigó, R. (2018). The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature Communications*, *9*(1). http://doi.org/10.1038/s41467-017-02772-x

Golden-Mason, L., & Rosen, H. R. (2017). Galectin-9: Diverse roles in hepatic immune homeostasis and inflammation. *Hepatology*. http://doi.org/10.1002/hep.29106

Goris, A., & Liston, A. (2012). The immunogenetic architecture of autoimmune disease. *Cold Spring Harbor Perspectives in Biology*, *4*(3). http://doi.org/10.1101/cshperspect.a007260

Greb, J. E., Goldminz, A. M., Elder, J. T., Lebwohl, M. G., Gladman, D. D., Wu, J. J., … Gottlieb, A. B. (2016). Psoriasis. *Nature Reviews Disease Primers*, *2*. http://doi.org/10.1038/nrdp.2016.82

Griffon, A., Barbier, Q., Dalino, J., Van Helden, J., Spicuglia, S., & Ballester, B. (2015). Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Research*, *43*(4). http://doi.org/10.1093/nar/gku1280

Gudjonsson, J. E., & Elder, J. T. (2007). Psoriasis: epidemiology. *Clinics in Dermatology*, *25*(6), 535–546. http://doi.org/10.1016/j.clindermatol.2007.08.007

Henderson, P., & Stevens, C. (2012). The Role of Autophagy in Crohn's Disease. *Cells*, *1*(4), 492–519. http://doi.org/10.3390/cells1030492

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. http://doi.org/10.1038/nprot.2008.211

Huang, Y., Xiao, S., & Jiang, Q. (2015). Role of rho kinase signal pathway in inflammatory bowel disease. *International Journal of Clinical and Experimental Medicine*, *8*(3), 3089–3097.

Hwang, S. T., Nijsten, T., & Elder, J. T. (2017). Recent Highlights in Psoriasis Research. *Journal of Investigative Dermatology*. http://doi.org/10.1016/j.jid.2016.11.007

Ivanov, A. I., Parkos, C. A., & Nusrat, A. (2010). Cytoskeletal Regulation of Epithelial Barrier Function During Inflammation. *The American Journal of Pathology*, *177*(2), 512–524. http://doi.org/10.2353/ajpath.2010.100168

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., & Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, *41*(1), 200–209. http://doi.org/10.1093/ije/dyr238

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830), 1497–1502. http://doi.org/10.1126/science.1141319

Kuleshov, M. V, Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90-7. http://doi.org/10.1093/nar/gkw377

Li, J., Wang, L., Jiang, T., Wang, J., Li, X., Liu, X., … Guo, M. (2016). eSNPO: An eQTL-based SNP Ontology and SNP functional enrichment analysis platform. *Scientific Reports*, *6*, 30595. http://doi.org/10.1038/srep30595

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. http://doi.org/10.1093/bioinformatics/btr260

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. http://doi.org/10.1016/j.cell.2008.03.029

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. http://doi.org/10.1038/nbt.1630

Menier, C., Riteau, B., Carosella, E. D., & Rouas-Freiss, N. (2002). MICA triggering signal for NK cell tumor lysis is counteracted by HLA-G1-mediated inhibitory signal. *International Journal of Cancer*, *100*(1), 63–70. http://doi.org/10.1002/ijc.10460

Miao, W., Masse, C., Greenwood, J., Kapeller, R., & Westlin, W. (2016). Potent and selective Tyk2 inhibitor highly efficacious in rodent models of inflammatory bowel disease and psoriasis. *Arthritis and Rheumatology*, *68*, 2415–2416. http://doi.org/10.1002/art.39977

Nestle, F. O., Kaplan, D. H., & Barker, J. (2009). Psoriasis. *The New England Journal of Medicine*, *361*, 496–509. http://doi.org/10.1525/jlin.2007.17.1.130.130

Polychronakos, C. (2011). Fine points in mapping autoimmunity. *Nature Genetics*. http://doi.org/10.1038/ng.1015

Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., … Chinnaiyan, A. M. (2010). HPeak: An HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, *11*. http://doi.org/10.1186/1471-2105-11-369

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., … Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, *4*(8), 651–657. http://doi.org/10.1038/nmeth1068

Schultz, B. M., Paduro, C. A., Salazar, G. A., Salazar-Echegarai, F. J., Sebastián, V. P., Riedel, C. A., … Bueno, S. M. (2017). A potential role of Salmonella infection in the onset of inflammatory bowel diseases. *Frontiers in Immunology*. http://doi.org/10.3389/fimmu.2017.00191

Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., … Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, *507*(7492), 371–375. http://doi.org/10.1038/nature13138

Song, L., & Crawford, G. E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *5*(2). http://doi.org/10.1101/pdb.prot5384

Strange, A., Capon, F., Spencer, C. C. A., Knight, J., Weale, M. E., Allen, M. H., … Trembath, R. C. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature Genetics*, *42*(11), 985–990. http://doi.org/10.1038/ng.694

Stuart, P. E., Nair, R. P., Ellinghaus, E., Ding, J., Tejasvi, T., Gudjonsson, J. E., … Elder, J. T. (2010). Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nature Genetics*, *42*(11), 1000–1004. http://doi.org/10.1038/ng.693

Stunnenberg, H. G., Consortium, T. I. H. E., Hirst, M., International Human Epigenome Consortium, & Hirst, M. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, *167*(5), 1145–1149. http://doi.org/10.1016/j.cell.2016.11.007

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a, … Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–50. http://doi.org/10.1073/pnas.0506580102

Torres, J. M., Gamazon, E. R., Parra, E. J., Below, J. E., Valladares-Salgado, A., Wacher, N., … Cox, N. J. (2014). Cross-tissue and tissue-specific eQTLs: Partitioning the heritability of a complex trait. *American Journal of Human Genetics*, *95*(5), 521–534. http://doi.org/10.1016/j.ajhg.2014.10.001

Tsoi, L. C., Spain, S. L., Ellinghaus, E., Stuart, P. E., Capon, F., Knight, J., … Elder, J. T. (2015). Enhanced meta-analysis and replication studies identify five new psoriasis susceptibility loci. *Nature Communications*, *6*. http://doi.org/10.1038/ncomms8001

Wang, J., Gamazon, E. R., Pierce, B. L., Stranger, B. E., Im, H. K., Gibbons, R. D., … Chen, L. S. (2016). Imputing Gene Expression in Uncollected Tissues Within and beyond GTEx. *American Journal of Human Genetics*, *98*(4), 697–708. http://doi.org/10.1016/j.ajhg.2016.02.020

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(D1). http://doi.org/10.1093/nar/gkt1229

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., … Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res*., gkv715-. http://doi.org/10.1093/nar/gkv715