Computational Study of Transcriptional Regulation - From Sequence To Expression

Shan Zhong

CMU-CB-13-101
May 2013

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Ziv Bar-Joseph, Chair
Roni Rosenfeld
Seyoung Kim
Takis Benos (University of Pittsburgh)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Keywords: Motif finding, Transcriptional regulatory network, p53, Protein binding microarray, Tissue specificity, EIN3, Ethylene response

To my parents, my wife, and my soon-to-be-born son.

Abstract

Transcription is the process during which RNA molecules are synthesized based on the DNAs in cells. Transcription leads to gene expression, and it is the first step in the flow of genetic information from DNA to proteins that carry out biological functions. Transcription is tightly regulated both spatially and temporally at multiple levels, so that the amount of mRNAs produced for different genes is controlled across different kinds of cells and tissues, as well as in different developmental stages and in response to different environmental stimulus. In eukaryotes, transcription is a complicated process and its regulation involves both *cis*-regulatory elements and *trans*-acting factors. By studying spatiotemporally what genes are regulated by which *cis*-elements and *trans*-factors, we can get a better understanding of how we develop, how we react to environmental signals, and the mechanisms behind diseases like cancer that, at least in part, result from failures in proper transcriptional regulation.

In this thesis, we present a suite of computational methods and analyses that, combined, provide a solution to problems related to the identification of DNA binding motifs, linking these motifs to the TFs that bind them and the genes that they control, and integrating these motifs and interactions with time series expression data to model dynamic regulatory networks. Specifically, we first develop a novel method for finding discriminative DNA motifs, motifs that are over-represented in a set of positive sequences but depleted in a set of negative sequences. Second, we present a new method of using protein binding microarray data combined with DNase I hypersensitivity and conservation data to predict tissue-specific transcription factor activities and binding sites. Finally, we extend the DREM framework which was previously developed by our group to study dynamic regulatory networks, and we use the improved version to analyze a biological dataset of gene responses in arabidopsis following ethylene treatment. Together, the methods and analyses presented contribute to the studying and understanding of transcriptional regulation.

Acknowledgments

Throughout my Ph.D. studies that led to this dissertation, I am thankful to many people that have kindly helped me along the way. I am fortunate enough to have Dr. Ziv Bar-Joseph as my Ph.D. advisor. Ziv's knowledge, vision and enthusiasm in computational biology have deeply impressed me over the years, and without his insightful guidance and strong support I would not have become what I am today. His optimism and encouragement have helped me grow my own confidence, and will keep affecting me for my life.

I would like to thank my committee members Seyoung Kim, Roni Rosenfeld and Takis Benos for providing many useful suggestions, comments and feedbacks on my thesis. I am particularly grateful to Guy Zinman, who was my office mate for 3 years and who had always been there to listen to and discuss with me about whatever problems I had met, and to Xin He and Marcel Schulz, whose wealth of knowledge and quick thinking had always brought new hopes when we faced difficulties in the projects we collaborated on. I would also like to thank all other people who had collaborated with me in different projects including Itamar Simon, Joseph Ecker, Dennis Kostka, Naftali Kaminski, Peter Huggins, Katherine Chang, Anthony Gitter, William Devanny, Idit Shiff, Rachel Beckerman, Oleg Laptenko, Carol Prives and Jason Ernst. Without them a lot of the works presented in this thesis would not have been complete. As a proud member in Ziv's Systems Biology group, I have benefited a lot from the sharing and discussion with all group members including Jason Ernst, Henry Lin, Peter Huggins, Marcel Schultz, Xin He, Saket Navlakha, Guy Zinman, Anthony Gitter, Hai-son Le, Aaron Wise and Siddhartha Jain. I would also like to thank Yuefeng Lin, Grace Huang, Ying Liu, Minli Xu, Armaghan Naik, Guozhen Zhang and Chi Song, from whom I had learned or was offered help in various aspects of my life during these years. In addition, I would like to thank the staffs at SCS help desk for their helps with problems I encounter with the Lane Center cluster, and Thom Gulish and Michelle Martin for their administrative assistance during my years in the program.

Last but not the least, I am deeply indebted to my parents Caiyun Ren and Huchen Zhong, my wife Hong Gu, and my soon-to-be-born son. It is only with your unconditional love and constant support that I am powered to overcome all difficulties in my life to pursue my dream.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and motivation

Transcription is the process during which RNA molecules are synthesized based on the information stored in DNAs in cells. Transcription leads to gene expression, and it is the first step in the flow of genetic information from DNA to proteins that carry out biological functions. In eukaryotes, three different types of RNA polymerases are involved in RNA synthesis, and protein-coding genes are transcribed by RNA polymerase II into messenger RNAs (mRNAs). Transcription is initiated by the binding of RNA polymerase II and a collection of proteins, called general transcription factors (TFs), to the core promoters (regions immediately upstream of the genes to be transcribed) to form the pre-initiation complex. Then in the elongation stage, the RNA polymerase unwinds the double strand DNA helix ahead, moves along the DNA template, and makes a complementary RNA molecule one base at a time. After the polymerase reaches certain signals in the DNA template to terminate transcription, it disassociates with the DNA template and the newly synthesized RNA is released.

Due to its importance, transcription is tightly regulated both spatially and temporally at multiple levels, so that the amount of mRNAs produced for different genes is controlled across different kinds of cells and tissues, as well as in different developmental stages and in response to different environmental stimulus. In eukaryotes, transcription is a complicated process and its regulation involves both *cis*-acting elements (DNA regions that are required for the proper expression of the nearby genes) and *trans*-acting factors (TFs that bind to the *cis*-acting elements to regulate the expression of other genes). In addition to the general TFs that bind to the core promoters of most genes to drive basal transcription, numerous specific TFs called activators (repressors) bind to *cis*-acting elements called transcription factor binding sites (TFBSs) around specific sets of genes. Their binding brings them into contact with the transcriptional machinery near the transcription start sites (TSSs) of these genes, and these TFs then increase (decrease) the expression of these genes by altering the rate at which they are transcribed. [2]

Transcriptional control is tightly linked to development. During animal development, a lot of the control of gene expression occurs at the transcription level. Several highly conserved families of TFs have been known to play important role in the regulation of development. For example, many Hox proteins, characterized by their homeobox domains, are important for forming the anterior-posterior patterning during development [136]; several Fox proteins, characterized by their forkhead domains, are involved in the development of liver, heart and many other tissues [92]. The determination of cell fates, the differentiation of cells into more specialized cell types and the change of cell morphologies during development are all under the delicate control of such TFs at precise times and locations.

Cells also respond to environmental stimuli through changes in their transcriptional programs. For example, from yeast to human, cells respond to a sudden increase in environmental temperature (heat shock response) by dramatically up-regulating the expression of Hsp (heat shock protein) genes under the control of TFs in the Hsf (heat shock factor) family [100], and in yeast this response also involves the activation of stress response TFs Msn2 and Msn4 [21]. The timely activation of such TFs in response to various environmental stress is vital to the survival and adaptability of the cells.

Problems in transcriptional regulation can have catastrophic effects in humans leading to many kinds of diseases. For example, dysregulation of genes that participate in cell cycle regulation, cell apoptosis, DNA replication and repair, and/or immunology have been frequently observed in different types of cancers including breast cancer [69], prostate cancer [86], gastric cancer [194], bladder cancer [141] and lung cancer [121], etc. In addition to cancer, dysregulation of genes have also been extensively observed in many other human diseases as well, including metabolic diseases like diabetes [135], cardiovascular diseases like dilated cardiomyopathy [13], autoimmune diseases like systemic lupus erythematosus [122], and neurological diseases like schizophrenia [197]. Such dysregulated genes are usually identified as differentially expressed between diseased and control samples in experiments that measure genome-wide expression levels using microarrays or more recently RNA sequencing. Their dysregulation in diseased samples usually results from errors in transcription programs that control their expression, like *trans*-mutations in the TFs themselves that regulate such genes, *cis*-mutations in their promoter regions that affect TF binding, or mutations higher up in the transcriptional cascade that affect these genes indirectly. Indeed, in several cases mutations in TFs have been shown to be associated with certain diseases. For instance, the transcription factor p53 is a famous tumor suppressor in human that maintains genome stability, and mutations in p53 have been observed across many types of cancers [120]. Mutations in the transcription factor Nkx2-5 have been shown to cause congenital heart disease [157]. Moreover, the roles of *cis*-mutations have been revealed by recent large-scale analysis [112], which reports that most of the single nucleotide polymorphism (SNP) variants that are detected to be associated with hundreds of human diseases locate in noncoding regions. Such noncoding variants are significantly enriched in DNase I hypersensitive sites (DHSs) that typically represent open chromatin regions bound by TFs, and the disease-associated variants often affect TF binding [112]. Therefore, transcriptional regulation also plays crucial role in maintaining our well-being and overall health.

In summary, by studying spatiotemporally what genes are regulated by which *cis*-elements and *trans*-factors, we can get a better understanding of development, how we react to environmental signals, and the mechanisms behind many diseases that result from failures in proper transcriptional regulation.

In this thesis, we present a suite of computational methods and analyses that, combined, provide a solution to problems related to the identification of DNA binding motifs, linking these motifs to the TFs that bind them and the genes that they control, and integrating these motifs and interactions with time series expression data to model dynamic regulatory networks. Together, the methods and analyses presented make valuable contributions to the studying and understanding of transcriptional regulation from a computational perspective.

1.2 High-throughput experimental methods that facilitate the study of transcriptional regulation

Many high-throughput experimental methods have been developed to study various aspects of transcriptional regulation either directly or indirectly. Below we provide a brief overview of several such methods that are relevant in various parts of this thesis.

1.2.1 Microarrays

Microarrays have been extensively used to measure the gene expression levels in cells in the last two decades[159]. There are two major types of microarrays, the cDNA microarray (or two-channel microararys) [154] and high-density oligonucleotide microarrays (or one-channel microarrays) [137]; the latter has become more popular and we'll focus on it here. In an oligonucleotide microarray experiment, oligonucleotides (probes) designed previously with known sequences, often from the genes whose expression levels are to be measured, are printed or synthesized *in situ* on the surface of a microarray chip using chemistry and photolithography. mRNAs are extracted from the samples, reverse transcribed into cDNAs, amplified, labeled, and added to the microarray for hybridization. The relative abundance of the mRNAs of the genes represented on the array are measured from their relative fluorescence intensities. Because the probes can be arranged on the microarray surface with a very high density, modern oligonucleotide microarrays can be used to measure the expression levels of all genes in typical organisms in a high throughput manner. In addition to measuring gene expression, specific types of microarrays have also been designed to detect SNPs [187] and copy number variations (CNVs) [140], investigate transcription factor binding preferences (see below), and detect genome-wide transcription factor binding locations following ChIP experiments (see below)

1.2.2 Protein binding microarray

Protein binding microarray (PBM) is a specific kind of microarray that allows the investigation of the binding specificities of a sequence-specific TF in a high-throughput and unbiased manner [17, 119]. In a typical PBM experiment, the microarray is composed of about $44,000$ cleverly designed probes of 60 nucleotide(nt) in length. Each probe contains a 24nt invariable primer sequence that attaches it to the microarray surface, and then a 36nt variable region designed using de Bruijn sequences of order 10, so that all possible 10-mers occur at least once on the array (Figure 1.1a,b). At the first step, universal primers labeled with Cy5 is added to the array, and then deoxyribonucleotides (dNTPs) labeled with Cy3 are added to synthesize double strand DNAs. The TF to be studied is expressed with a glutathione S-transferase (GST) tag, purified and then applied to the microarray. After this, fluorophore-conjugated antibody to GST is applied to the array, and the binding strength of the TF to each probe is measured by fluorescent intensities [17] (Figure 1.1c). Unlike ChIP-based techniques that study the *in vivo* genome-wide binding locations of a TF (see below), PBM has the advantage that it does not require specific antibody for the TF of interest, and it does not require knowledge about the genome of the species from which the TF comes. PBM has been used to reveal the binding profiles of hundreds of TFs in yeast [207], worm [60], mouse [6] and arabidopsis [24].

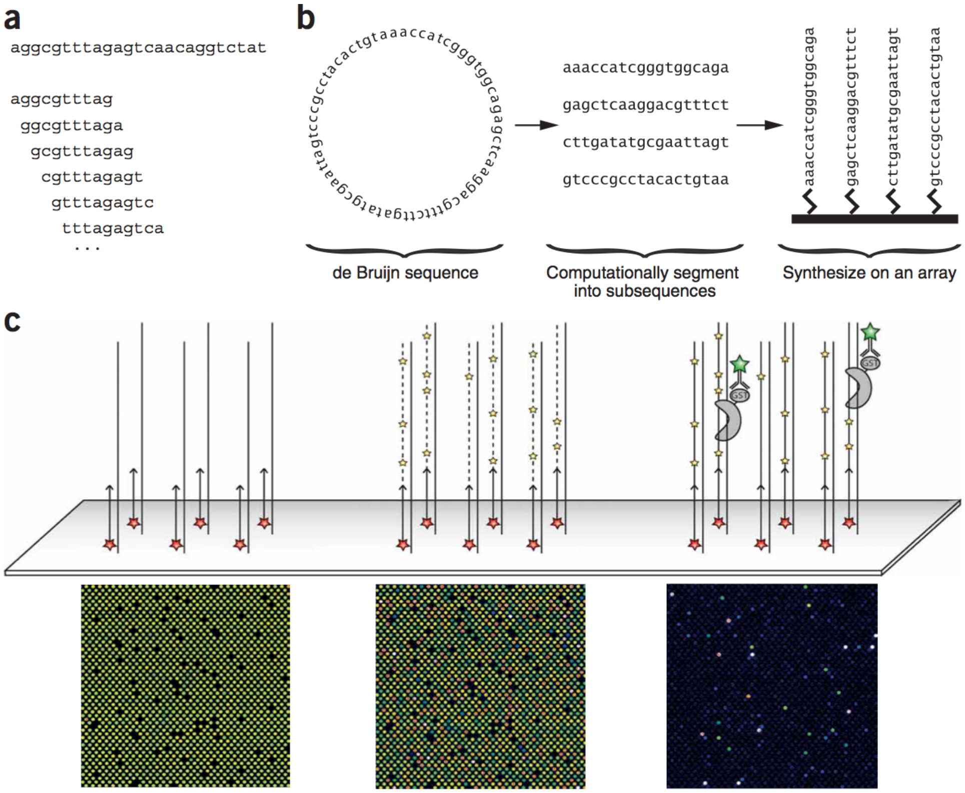


Figure 1.1: Protein binding microarray experiment. (a,b) Illustration of probes on a hypothetical microarray with de Bruijn sequence of order 3, so that all 3-mers appear at least once on the array surface (in practice an order 10 is used). (c) PBM experimental procedure. See text for details. This figure is from [17].

PBM allows the elucidation of the sequence-specific binding preferences of invididual TFs to an unprecedented scale and subtlety. However, it also poses computational problems in the process of converting the intensity readouts to biological knowledge. Specifically, since the length of a TFBS is typically 8-10bps, effective methods are needed for obtaining the real binding profiles of a TF from the intensity measurements on the 36nt probes. Several methods have been proposed for this purpose [1, 5, 17, 204, 207]. In this thesis, we develop a novel method, PLAR-PBM, for using PBM data to infer TF binding profiles, and we show that our method outperforms the other methods mentioned above in predicting *in vivo* TF binding sites (Chapter 3).

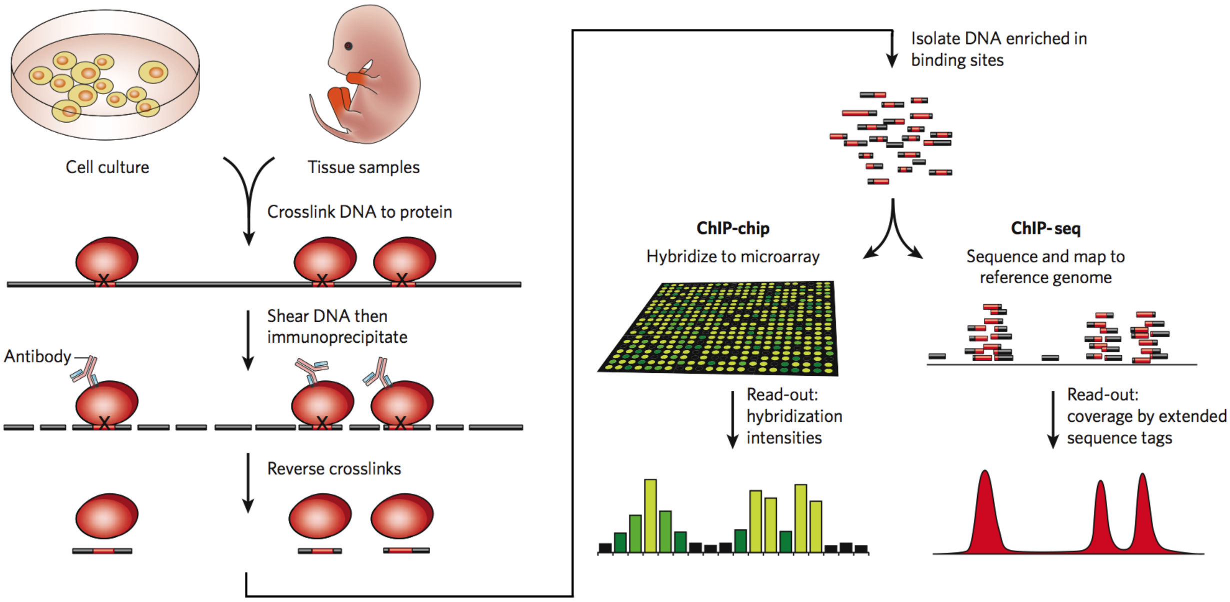
1.2.3 RNA sequencing

In recent years, with the fast development and reduced cost of sequencing technology, RNA sequencing [188] is becoming a popular method for measuring gene expression levels. In a typical RNA-seq experiment, RNAs extracted from samples are first converted to cDNA fragment library and amplified, and then specific sequencing adaptors are ligated to the ends of the cDNA fragments in the library. Each fragment in the library is read by sequencers and a short read is output. These reads are then aligned to the reference genome, generating a high-resolution map of expression levels at each position in the genome. Compared with microarrays, RNA-seq has the advantage that it has higher resolution and larger dynamic range, and it allows the detection of novel gene fusion and alternative splicing isoforms. In this thesis, RNA-seq experiments were used to measure gene expression profiles in arabidopsis following ethylene treatment (Chapter 4).

1.2.4 ChIP-chip and ChIP-seq experiments

Chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-chip) [23] or sequencing (ChIP-seq) [134] has been developed to study genome-wide TF binding *in vivo* (Figure 1.2). In such experiments, the *in vivo* protein-DNA interactions are first cross linked by formaldehyde, and then these cross linked chromatin is sheared into fragments. The TF of interest is immunoprecipitated with specific antibody, and then the cross linking is reversed to release the bound DNA fragments. The location of these DNA fragments bound by the TF is then determined by either hybridization to specific microarray containing promoter regions from the genome (ChIP-chip), or by direct sequencing and aligning to the reference genome computationally (ChIP-seq). In this thesis, ChIP-chip experiments were performed to detect binding locations of different forms of human p53 (Chapter 2), and ChIP-seq experiments were performed to identify binding locations of EIN3 in arabidopsis at different time points following ethylene treatment (Chapter 4). Moreover, existing ChIP-seq datasets for several TFs were used to demonstrate the power of

DECOD, a new discriminative motif finding method that we developed (Chapter 2), and to evaluate the performance of several methods that analyze PBM data, including ours, for classifying known *in vivo* TF binding sites (Chapter 3).



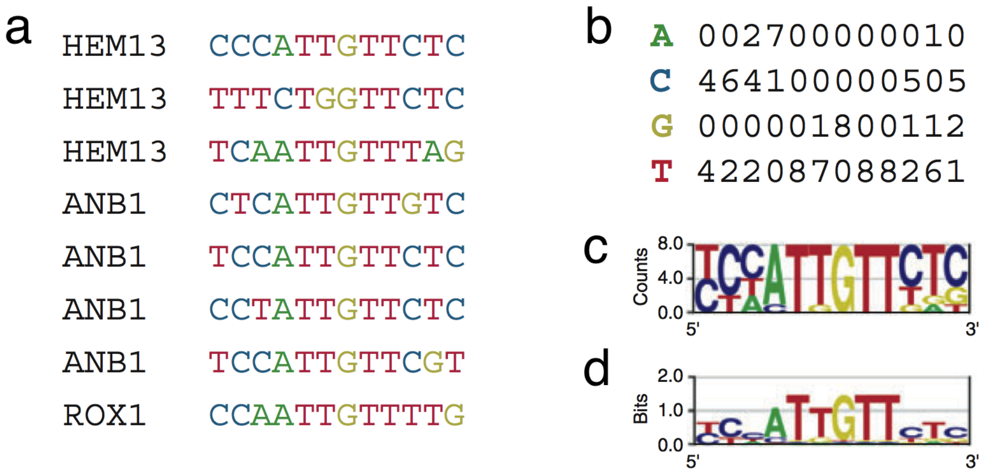
Cell culture
Tissue samples
Crosslink DNA to protein
Shear DNA then immunoprecipitate
Antibody
Reverse crosslinks
Isolate DNA enriched in binding sites
ChIP-chip
Hybridize to microarray
ChIP-seq
Sequence and map to reference genome
Read-out: hybridization intensities
Read-out: coverage by extended sequence tags

Figure 1.2: ChIP-chip and ChIP-seq experiments. See text for details. This figure is modified based on [185].

1.2.5 Motif and position weight matrices

Binding sites in the genome for the same TF are usually not exactly the same; instead, they are often rigid at certain positions and flexible at others. These binding sites are often referred to as occurrences of the *motif* for the corresponding TF. Such motifs are usually represented by a position weight matrix (PWM; also called position-specific scoring matrix, PSSMs) to provide a simple, intuitive and also informative view of the TF's binding preference. Traditionally, the PWM for a TF is constructed by first aligning a group of experimentally determined known binding sites for the TF (Figure 1.3a). After the alignment, a count matrix can be generated by counting the number of times each nucleotide is used at each position in the alignment (Figure

1.3b), and this count matrix is then converted to a frequency matrix where the counts for each nucleotide at each position are replaced by their probabilities at that position. Sequence logos [156] are frequently used to visualize such count or frequency matrices, where at each position, the relative height of each nucleotide reflects their relative frequencies (Figure 1.3c), and the total height of each position reflects the information content at that position (Figure 1.3d). PWMs are then generated by converting the nucleotide frequencies to log likelihood or log-odd ratios which takes into account a background model, so that a new subsequence can be scored by summing up the corresponding elements in the PWM. See [40] for a thorough introduction.



HEM13	CCCATTGTTCTC
HEM13	TTTCTGGTTCTC
HEM13	TCAATTGTTTAG
ANB1	CTCATTGTTGTC
ANB1	TCCATTGTTCTC
ANB1	CCTATTGTTCTC
ANB1	TCCATTGTTCGT
ROX1	CCAATTGTTTTG

A	002700000010
C	464100000505
G	000001800112
T	422087088261

Figure 1.3: PWM model for TF motif. (a) A group of aligned short sequences known to be bound by the TF. (b) Count matrix generated from (a). (c,d) Sequence logos generated from the count matrix. See text for details. This figure is modified based on [40].

One of the major drawbacks of the PWM motif model is that it assumes independence between positions, which is not always true in reality [6, 105]. Recently, more complicated models have been introduced to address this [163, 166, 206], but PWM is still most popularly used due to its simplicity. In this thesis, we use PWM models in DECOD for discriminative motif finding (Chapter 2), and later in PLAR-PBM, we show that by using a k-mer based model that does not assume independence between positions to represent TF binding profiles, we indeed achieve better prediction accuracy when classifying real TF binding sites (Chapter 3).

1.2.6 DNase I hypersensitivity and DNase-seq

DNA *in vivo* are not naked but instead wrapped around by various histone proteins forming nucleosomes, and these nucleosomes are the basic unit of the chromatin. For a TF to bind to a segment of DNA *in vivo*, the DNA need to be exposed from nucleosome protection and become accessible to the TF. Early studies showed that genomic regions that are sensitive to DNase I cleavage, called DNase I hypersensitive sites (DHSs), correlate with open chromatins that are accessible to TF binding [59]. Therefore, the incorporation of such DHS information can help reduce false positives in predicting genome-wide TF binding sites. Recently, with the advancement of sequencing technology, DNase-seq has been developed to measure genome-wide *in vivo* DHSs in a high throughput manner [35]. Briefly, nuclei are extracted from the cells under investigation and subjected to DNase I cleavage. The cleaved segments are sequenced and mapped to the reference genome, and DHSs are revealed as regions that have a high number of reads mapped. The ENCODE project has produced such DHS data for many tissues and cell types in human and mouse [183]. In this thesis, we use such data for 55 mouse tissue/cell types and combine them with PBM data to predict TF activities and binding sites (Chapter 3).

1.3 Overview of the thesis

This thesis is composed of three parts that together present novel computational approaches to study various aspects of transcriptional regulation. Below we provide a brief overview of the thesis.

In Chapter 2, we present DECOD, a new tool for discriminative motif finding. DECOD allows the finding of motifs that are over-represented in one set of sequences and depleted in another set. One unique feature of DECOD is that its running time does not depend on the size of the input sequences, and therefore it can be run on input sets that contain thousands of or even more sequences to discover discriminative motifs. Input of this scale is typical in the current era where large amount of data is generated by high-throughput sequencing methods such as in ChIP-seq. We demonstrate that DECOD is superior in terms of both speed and accuracy as compared with many other existing tools for discriminative motif finding. DECOD is written in Java, and therefore can be run on different platforms. It comes with a friendly easy-to-use GUI interface that is convenient for biologists to use. This chapter is based on our paper published [70].

In Chapter 3, we develop a new method, PLAR-PBM, for analyzing protein binding microarray data. We also present an integrated model that combines PBM data with DNase I hypersensitivity data to predict tissue-specific TF binding. PLAR-PBM uses a biophysically-motivated k-mer based model to infer binding preferences of TFs from PBM data, and it outperforms several existing methods for the same purpose when evaluated on *in vivo* data. When combined with DNase data, we were able to predict tissue-specific TF activities and binding sites with high accuracy. This allowed us to generate a resource for computationally predicted tissue-specific TF binding sites for 284 TFs across 55 mouse tissue/cell types. Such a resource can be very useful to biologists studying transcriptional regulatory network. This chapter is based on our submitted paper [205].

In Chapter 4, we extend DREM [46], a tool that studies dynamic transcriptional regulatory network by integrating expression with binding data previously developed by our group, to allow the use of dynamic binding data and the integration with DECOD [70] to perform discriminative motif finding. These, among other new features, have been incorporated into a new version of DREM 2.0. Using DREM 2.0, we analyzed gene responses in arabidopsis following ethylene treatment, and showed that ethylene triggers four waves of gene transcriptions under the master regulator EIN3. This chapter is based on our published paper for DREM 2.0 [158] and on the submitted paper for the analyses in arabidopsis [24].

Chapter 2

DECOD: Fast and accurate discriminative motif finding

2.1 Introduction

DNA motif discovery has been a central problem in computational biology for almost two decades. Many methods based on word enumeration or probabilistic models including position weight matrices (PWMs) and Hidden Markov models (HMMs) have been developed for this task [38]. Word enumeration-based methods are usually only able to find short motifs and tend to fail when the motif includes weak positions [38]. Most probabilistic methods involve iteratively scanning the input sequences to identify potential motifs and then updating the motifs to improve the likelihood of the model until convergence [8, 52, 87, 101, 151, 170, 181, 196]. In such methods, motifs are usually defined as subsequences, which are present at a much higher rate than expected when compared with a background model [40].

The use of motif discovery methods has dramatically increased over the last few years due to the rise in sequencing capacity and the advancement of other high-throughput methods. These methods are routinely used to identify and predict transcription factor binding sites [19, 68], protein phosphorylation sites [148, 160], microRNA targets [72, 98] and alternative splicing locations [179]. However, these high-throughput methods have also led to new requirements from motif search algorithms. The first is speed. Many studies now routinely search for motifs in very large sets of input sequences. For example, several ChIP-Seq experiments identify thousands of targets for specific mammalian transcription factors [27, 74, 149, 199]. The second requirement is for identifying discriminative motifs [168]. Unlike traditional motif searches that are performed against a general background model, in discriminative motif search one looks for motifs that are present at a high rate in a positive set compared to a negative set. These sets can be genes that are up- or downregulated at a specific time point or condition [46], proteins that are initially co-localized but later diverge [97], genes that are bound in one condition by a TF but not in another [63], etc. These and other studies, including cross species analysis and methods for modeling gene regulation, require discriminative motif discovery methods that can scale to large datasets.

Several discriminative motif-finding methods have been developed so far. DIPS [169] uses a probabilistic score to quantify the difference in the number of occurrences of a PWM between two sets of sequences and uses heuristic hill climbing to search the sequences for motifs that maximizes this score. ALSE [93] uses a target function based on the hypergeometric distribution. This function searches for a PWM using an EM-like heuristic and then evaluates the likelihood that the PWM it identified represents a real motif. DEME [143] performs a combination of global and local search to find a PWM that maximizes the conditional log likelihood of the sequence labels given the sequences and models parameters. Seeder [49] is a word-based enumerative method. It first generates seeds by finding significantly enriched words in the positive set based on a word-specific background probability distribution, and then iteratively extends these seeds to form a new PWM and updates the seeds until convergence. CMF [109] is also a word-based method that starts by finding enriched words in the positive set based on a z-score, and then iteratively updates the motif model and rescans the sequences to update the seeds and avoid false positives until convergence. See the next section for a more detailed description of these existing methods.

While the above methods can successfully identify discriminative motifs, they usually do not scale well for large sequence datasets since they are based on repeated analysis of the positive and negative sequences. For example, DIPS [169] was only suggested to be run on tens of sequences with length around 1000bp, and even so its run time is very long (several hours). The running time of DEME [143] depends quadratically on the size of the positive sequences, making it prohibitive for most motif discovery tasks. Other methods are also slow when dealing with large datasets as we show in Results.

DME [172] attempts to address the speed issue by enumerating over a discrete space of pre-defined matrices representing possible motifs. It then uses a log likelihood ratio as a target function to score the overrepresentation of a motif matrix in one set of sequences versus another. However, while DME is indeed very fast, it is based on a pre-defined set of matrices and is thus often restricted in terms of the set of motifs it can identify. In addition, DME ignores the context information encoded as part of the sequences, which may lead to a shifted PWM that does not accurately represent the real motif.

In this chapter, we present a new method that addresses both the speed and accuracy issues for discriminative motif finding. Our method, *deco*nvolved *d*iscriminative motif finder (DECOD), only uses k-mer counts and so does not depend on the size of the input set. To compensate for the errors introduced from ignoring the dependence between the consecutive and overlapping k-mers in the sequences that they are from (the context of a k-mer), we use a deconvolution method that accounts for the higher rates of k-mers containing subsets of the true motif. We applied the method to simulated and biological benchmark data and compared it with previous methods. As we show, our method enables motif discovery in cases that could not have been studied before due to the size of the input, and it outperformed other methods in terms of both accuracy and running time. We used our method to study various post- translational modification of the human transcription factor p53. We performed new ChIP-chip experiments and identified different sets of binding targets for the p53 mutants. Using our new motif discovery algorithm we were able to identify a number of potential co-factors of p53 and study the way in which they interact with p53.

2.2 Existing methods for discriminative motif finding

In contrast to traditional motif finding methods, discriminative motif finding methods requires a negative set of sequences to be supplied and compared against. Several discriminative motif finding methods have been developed. Below we provide a detailed description of the methods that we compared in this work, highlighting the similarities and differences between DECOD and these methods when possible.

2.2.1 DME (*d*iscriminative *m*atrix *e*numerator)

Assuming that each k-mer in the sequences is either motif or background, DME [172] uses a likelihood model to score for motif overrepresentation in the positive sequences relative to background sequences. It aims to maximize a target function that represents a modified version of the log ratio of the likelihood of the motif and background models given the positive set to that of the motif and background models given the negative set. And DME uses exhaustive search to find the motif model that maximizes this target function. To improve search efficiency, DME first only searches over a very sparse discrete PWM space in which the columns of the PWMs are only of several representative types. Then DME uses a refinement step to extend the search by including matrices in the neighborhood of the matrix found in the global search. Moreover, instead of using the original log likelihood ratio as the target function, DME uses a modified version which can be calculated very fast with the assumptions that the base frequencies in the positive and negative sequences are close, and that the motif occurrences in the positive sequences are not dense. Our method is similar to DME in that both assumes that each k-mer comes from either a motif or a 0th-order background model, and that both employ global and local searches to improve search speed. However, our method explicitly assumes the probability of the motif and background models being used while DME does not model these. Moreover, our target function is not based on likelihood models but instead based on the expected number of times that the PWM is used in generating the sequences. Furthermore, our method does not make the assumption as made by DME that the base frequencies are similar in the positive and negative sequences, and our method uses deconvolution to take into account the k-mer contexts that DME ignores. Also our method does not search over the PWM space directly but instead searches over the k-mers from which the PWM model is constructed.

2.2.2 DEME (*d*iscriminatively *e*nhanced *m*otif *e*licitation)

DEME [143] also uses a probabilistic approach to model the sequences. Given labeled sequences, DEME aims to find a set of parameters for the data model (including the motif model, the background model, the probability of a positive sequence containing a motif and the prior probability of a sequence being labeled positive), that maximizes a target function describing the conditional log likelihood of the sequence labels given the sequences themselves and the above parameters. DEME also uses a combination of global and local searches in the optimization process. In global search, DEME performs substring search and branch search to find strings from positive sequences, allowing mutations, whose corresponding motif model has the best objective function score. It then uses conjugate gradient to perform local search to further refine the model parameters. One unique feature of DEME is that it is able to work on protein sequences, and it can incorporate prior knowledge about protein residue characteristics by using a Bayesian prior on motif columns. DECOD is similar to DEME in that both uses 0th order background model, and both involves global and local searches in the optimization of the target function. DECOD takes the probability of motif occurrence in positive sequences as a user-input parameter (by default assuming it to be once per positive sequence), while DEME tries to learn it automatically.

Notably, DEME can only find one discriminative motif from an input dataset, whereas DECOD is able to probabilistically remove the signals of a previously found discriminative motif in order to find the next one.

2.2.3 DIPS (*d*iscriminative *P*WM *s*earch)

DIPS [169] assumes that sequences are generated by a 0th order HMM, and it uses a probabilistic score (w-score) to count the occurrences of a PWM in each sequence, which accounts for both the number and strengths of motif occurrences. The w-score of a PWM in a sequence is the sum of the number of times that the PWM is used in all possible parses of the sequence in the HMM model, weighted by the probability of each parse. DIPS then uses the difference in the average w-score of a PWM between sets of positive and negative sequences as the target function. DIPS employs heuristic hill climbing to search for a PWM that maximizes the target function. Our method is very similar to DIPS since both model the sequences by 0th order HMMs, and both aim to maximize the differences in the expected number of times that a PWM is used in generating the positive and negative sequences respectively. Also the search strategy that we used is inspired by DIPS. However, the actual target functions used by the two methods are different. Our method does not calculate the w-scores of a PWM by working on the sequences and calculating the probability of each possible parse according to the HMM model. Instead we work on the k-mers extracted from the sequences directly. This makes DECOD much faster than DIPS. Moreover, in the search process we both reduce the search space and reduce the amount of computation involved in calculating the target function score in order to speed up the optimization process particularly for longer motifs, while DIPS did not make these attempts.

2.2.4 CMF (*c*ontrast *m*otif *f*inder)

CMF [109] is a word-enumeration based method for discriminative motif finding. Given sets of positive and negative sequences, CMF first calculates a z-score of all k-mers to find the k-

mers and its neighborhood that are most enriched in the positive compared to the negative set. It then uses the k-mers found to create two count matrices in the positive and negative sequences (representing false positives) respectively to be used as seed, and a PWM is generated by taking their differences. After that, CMF scans k-mers in all sequences in the positive and negative set using the PWM and a 1st-order Markov background model, and calculate a likelihood ratio score for each k-mer. It then applies a threshold on the likelihood ratios by controlling FDR, and all k-mers that pass the threshold are used to create a new PWM. The process is iterated until convergence.

2.2.5 ALSE (*al*l *se*quences)

ALSE [93] aims to find a motif model that maximizes a target function that represents the likelihood of the motif being the true one given the input positive and negative sequences. The target function is calculated based on hypergeometric distribution. ALSE first finds a set of seed matrices using a Voting algorithm, and then iteratively refines the matrices in EM-like iterations until no further improvement can be made in the likelihood function.

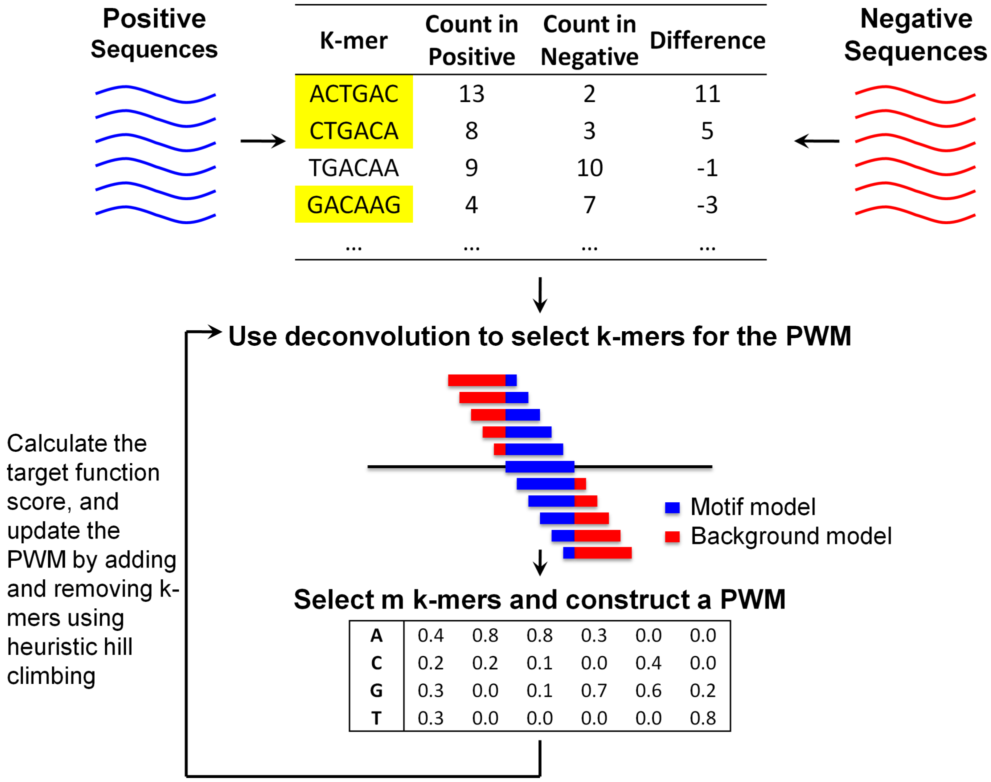
2.2.6 Seeder

Seeder [49] is a word-enumeration based method that starts by enumerating all k-mers of a given seed length. For each k-mer, the Hamming distance between the k-mer and its best matching subsequence (called “substring minimal distance”, SMD) in the positive and negative sets are then calculated respectively. The latter is used to calculate a word-specific background probability distribution, which is in turn used to evaluate the significance of the enrichment of each k-mer in the positive sequences based on the sum of its SMDs to positive sequences. Seed PWMs are generated from matches to the most enriched k-mers in each positive sequence, and the seeds are iteratively extended to form new PWMs of the desired motif width, and then the seeds are updated. The entire process is repeated until convergence.

2.3 Method: The DECOnvolved Discriminative motif discovery method

Similar to other methods [49, 93, 143, 169, 172], DECOD starts with a user-specified motif length k. Given k, we extract all k-mers from the positive and negative sequences (Figure 2.1). Following this step the entire analysis is only performed on the k-mer counts table. Since the size of this table is independent of the number and length of the input sequences, DECOD scales very well to large datasets.

We assume a generative mixture model for k-mer distributions: Each k-mer is either generated by the motif model represented by a PWM, or by the background model (similar to a zeroth-order HMM). Following [169], DECOD searches for a PWM that maximizes a discriminative target function: the difference in the expected number of times that the motif model is used to generate the positive and negative sequences (Figure 2.1, top). The PWM is constructed from a subset of the k-mers which are selected based on the k-mer count table (termed "the site set" [169], highlighted k-mers in Figure 2.1). While using only the k-mer counts provides significant speed benefits with large input datasets, such representation ignores important context information for each k-mer within a sequence. This may result in selecting shifted versions of the same k-mers that lead to a convolved (and inaccurate) PWM (Figure 2.1, middle). To correct for this we use a deconvolution method that accounts for the higher rates of k-mers that contain a subset of the true motif in the positive set. In an iterative process we continuously improve our PWM by adding and removing k-mers from the site set using heuristic hill climbing search methods until convergence. Once the algorithm converges we remove instances of the identified PWM from the k-mer count table, and then search for a second PWM and so forth. We discuss each of these steps in details below.


Positive Sequences
Negative Sequences
Use deconvolution to select k-mers for the PWM
Calculate the target function score, and update the PWM by adding and removing k-mers using heuristic hill climbing
🟦 Motif model
🟥 Background model
Select m k-mers and construct a PWM

K-mer	Count in Positive	Count in Negative	Difference
ACTGAC	13	2	11
CTGACA	8	3	5
TGACAA	9	10	-1
GACAAG	4	7	-3
...	...	...	...

A	0.4	0.8	0.8	0.3	0.0	0.0
C	0.2	0.2	0.1	0.0	0.4	0.0
G	0.3	0.0	0.1	0.7	0.6	0.2
T	0.3	0.0	0.0	0.0	0.0	0.8

Figure 2.1: Overview of DECOD. We extract counts of all k-mers in the positive and negative sequences (top) and store them in a k-mer count table. Next, we search for a discriminative PWM that matches many k-mers on the positive set while only matching a few on the negative set. The PWM is constructed using a site set containing a small number of k-mers (highlighted in yellow). To determine which k-mers to include in the site set we use a deconvolution based target function (middle) which overcomes the lack of context information for the k-mers. Once appropriate k-mers are identified we revise the PWM (bottom) and the process is repeated until no further improvement to the target function can be achieved.

2.3.1 The mixture model for k-mers

DECOD uses the following mixture model that includes a motif component **Z** and a background component **B** to model the k-mer distribution **M**:

$$\mathbf{M} = p\mathbf{Z} + (1-p)\mathbf{B} \tag{2.1}$$

Here, $\mathbf{Z}$ and $\mathbf{B}$ are the probability distributions over the k-mers (i.e. non-negative vectors of dimension 4^k whose entries sum to 1) generated by the motif and background models respectively, and p is the probability of motif occurrence. The mixture model $\mathbf{M}$ can also be considered as a zeroth order HMM that generates k-mers as follows: (i) choose a hidden state h from $\{z, b\}$ with state probabilities p and $1-p$ respectively; (ii) if $h = z$, emit a k-mer according to the distribution $\mathbf{Z}$; if $h = b$, emit a k-mer according to the distribution $\mathbf{B}$.

2.3.2 The motif component Z and deconvolution

The simplest way to model $\mathbf{Z}$ by a PWM θ is to define each element $\mathbf{Z}_a$ to be

$$\mathbf{Z}_a = \Pr(a|\theta) = \prod_{i=1}^{k} \theta_{i,a_i} \equiv \theta^a \tag{2.2}$$

in which $a = a_1 \dots a_k$ is a k-mer, θ_{i,a_i} is the entry for the letter a_i in the i's column of θ and we use θ^a as a shorthand notation for $\Pr(a|\theta)$. We call such $\mathbf{Z}$ *simple motif component*.

However, our method for extracting overlapping k-mers, while greatly speeding up computational time for large input datasets, ignores the context of the k-mers. Thus, several k-mers that do not fully match the motif may still overlap parts of it and thus may be overrepresented in the data. To overcome this, note that each k-mer in its context can be generated by $2k-1$ combinations of the motif component and the background component (Figure 2.1). Thus instead of the simple PWM mixture component, we define the following *convolved motif component*:

$$(2k-1)\mathbf{Z}_{\text{convolved}} = \mathbf{Z}_{-(k-1)} + \cdots + \mathbf{Z}_0 + \cdots + \mathbf{Z}_{k-1}, \tag{2.3}$$

where $\mathbf{Z}_0$ is the k-mer frequencies obtained from the PWM θ, and $\mathbf{Z}_j$ the k-mer frequencies from a PWM obtained by taking the first j columns of θ (or the last j columns if $j < 0$), and adding $k - j$ columns of background as a prefix (or suffix if $j < 0$). Note that using the convolved motif component, the mixture model becomes

$$\mathbf{M} = p(2k-1)\mathbf{Z}_{\text{convolved}} + \left[1-(2k-1)p\right]\mathbf{B} \tag{2.4}$$

2.3.3 Target function of the discriminative PWM search

We are given a set of positive sequences S_+ and a set of negative sequences S_- as input. Normalized k-mer counts are extracted and denoted by X for the positive set and Y for the negative set, and together they form the input for DECOD. Assuming that X was generated by the mixture model, the expected number of times that the motif component $\mathbf{Z}$ was used in the zeroth-order HMM is

$$w(X;\mathbf{Z}) = \sum_{a\in\Sigma^k}\left(\frac{p\mathbf{Z}_a}{p\mathbf{Z}_a+(1-p)\mathbf{B}_a}\right)\cdot X_a \tag{2.5}$$

in which $\mathbf{Z}_a = \Pr(a|\theta)$ is the probability of observing a under the motif model, $\mathbf{B}_a = \Pr(a|B)$ is the probability of observing a under the background model, and X_a is the count of a in the positive sequences. A similar expression can be written for Y. Following [169], given X, Y as input, we aim to maximize the expected difference

$$F(Z) = w(X;\mathbf{Z}) - w(Y;\mathbf{Z}) = \sum_{a\in\Sigma^k}\left(\frac{pZ_a}{pZ_a+(1-p)B_a}\right)\cdot(X_a - Y_a) \tag{2.6}$$

in which Z and B represent the estimated distribution on k-mers as discussed above. The background B is estimated from the base frequencies of the input sequences using a simple zeroth-order model. Below we will regard B as a PWM as well, with all columns being equal.

Assuming a simple motif component $\mathbf{Z}$, let θ denote the PWM for $\mathbf{Z}$. Then the discriminative score can be written as

$$F(\theta) := F(Z(\theta)) = \sum_{a\in\Sigma^k}(X_a - Y_a)\frac{p\theta^a}{p\theta^a+(1-p)B^a} \tag{2.7}$$

For a convolved motif component $\mathbf{Z}$, a similar formula can be derived. For PWMs A, B of length k, let $[\underline{A}_i\overline{B}_{k-i}]$ denote the PWM obtained by concatenating the last i columns from A with the first $k-i$ columns from B. Then the discriminative score for the convolved motif component is:

$$F(\theta) := F(Z(\theta)) = \sum_{a \in \Sigma^k} (X_a - Y_a) \cdot \frac{p\left[\theta^a + \sum_{j=1}^{k-1}\left([\underline{\theta}_j \overline{B}_{k-j}]^a + [\underline{B}_j \overline{\theta}_{k-j}]^a\right)\right]}{p\left[\theta^a + \sum_{j=1}^{k-1}\left([\underline{\theta}_j \overline{B}_{k-j}]^a + [\underline{B}_j \overline{\theta}_{k-j}]^a\right)\right] + [1-(2k-1)p]\, B^a} \tag{2.8}$$

As before, our aim is to find a PWM θ that maximizes the above function. The details of the search procedure is described in Section 2.3.4.

In practice, when the two input datasets are not equal in size and have different base frequencies, we replace the counts X_a and Y_a above with the frequencies of the k-mer a in the two sets, and we use different Bs estimated from the two sets respectively, and calculate $w(X; \mathbf{Z})$ and $w(Y; \mathbf{Z})$ separately. Also for the probability of motif occurrence p, we show that similar to DIPS [169], our method is not sensitive to the choice of this parameter (Section 2.6.5), and we set it to be once per positive sequence.

2.3.4 Searching for the discriminative PWM that optimizes the target function

We adopt a discretized heuristic hill climbing approach very similar to DIPS [169] to search for the PWM that optimizes Equation 2.8. The search space for θ is restricted to empirical PWMs of the form $\theta(T)$, where T is a subset of m k-mers in S_+ called *site set* [169]. The subset size m is a parameter called *motif cardinality* [169].

Heuristic hill climbing is used to find a local subset T that maximizes $F(\theta(T))$. Each hill climbing step of composed of (i) delete: remove one k-mer from T that contributes the least to the score, and (ii) add: add one k-mer from $S_+ \backslash T$ to T that contributes the most to the score. In the delete step, every possible $t \in T$ is tested to find the t_i that maximizes $F(\theta(T \backslash t_i))$. Then T is updated by setting $T \leftarrow T \backslash t_i$. This step is fast since the size of the PWM set is small (usually

20). In the add step, every k-mer $s \in S_+ \backslash T$ is considered for being added to T. This step is slow since it requires us to loop over all k-mers. To make the calculations faster, partial derivatives are used to estimate $F(\theta(T \cup s))$ as follows:

$$F(\theta(T \cup s)) - F(\theta(T)) \approx \nabla F(\theta(T)) \cdot \delta \tag{2.9}$$

where $\delta = \theta(T \cup s) - \theta(T)$. Detailed derivation of the partial derivatives is provided below. Choices of s are sorted according to their estimated values of $F(\theta(T \cup s))$. Then $F(\theta(T \cup s))$ is computed exactly for each choice of s in sorted order until an s is found that satisfies $F(\theta(T \cup s)) > F(\theta(T))$, and T is updated by setting $T \leftarrow T \cup s$. In practice, we terminate the hill climbing search if the top 500 k-mers on the ranked list do not lead to an improvement to the discriminative score.

2.3.5 Computation of partial derivatives

For a simple motif component $\mathbf{Z}$, partial derivatives of F can be written succinctly:

$$\frac{\partial F}{\partial \theta_{ij}} = p \cdot (1-p) \cdot \sum_{a \in \Sigma^k} (X_a - Y_a) \frac{a_{ij} \theta_{ij}^{-1} \theta^a B^a}{(p\theta^a + (1-p)B^a)^2} \tag{2.10}$$

Here the k-mer a is represented as a $4 \times k$ matrix in which each element $a_{ij} \in \{0, 1\}$ and the columns sum to 1.

For a convolved motif component, partial derivatives of F is:

$$\frac{\partial F(\theta)}{\partial \theta_{mn}} = p \cdot [1-(2k-1)p] \cdot \sum_{a \in \Sigma^k} c(a) \frac{B^a}{(pA + [1-(2k-1)p]B^a)^2} \cdot \frac{\partial A}{\partial \theta_{mn}} \tag{2.11}$$

in which

$$\begin{aligned}\frac{\partial A}{\partial \theta_{mn}} = &\prod_{\substack{i=1\\ i\neq n}}^{k} \left(\sum_{j=1}^{4} \theta_{ji} a_{ji} \right) \cdot a_{mn} + \\ &\sum_{l=1}^{k-n} \left[\prod_{i=1}^{l} \left(\sum_{j=1}^{4} b_j a_{ji} \right) \cdot \prod_{\substack{i=l+1\\ i\neq l+n}}^{k} \left(\sum_{j=1}^{4} \theta_{j,i-l} a_{ji} \right) \cdot a_{m,l+n} \right] + \\ &\sum_{l=k-n+1}^{k-1} \left[\prod_{i=l+1}^{k} \left(\sum_{j=1}^{4} b_j a_{ji} \right) \cdot \prod_{\substack{i=1\\ i\neq n-k+l}}^{l} \left(\sum_{j=1}^{4} \theta_{j,k-l+i} a_{ji} \right) \cdot a_{m,n-k+l} \right]\end{aligned} \tag{2.12}$$

Detailed derivation of the above is in Appendix A.

2.3.6 Identifying multiple PWMs representing combinatorial regulation

After a PWM is found, if desired, we remove the signals of that PWM from the k-mer count table and start searching for a second PWM. Our hill climbing algorithm assumes (an estimate of) the plant probability p is given. Initially p is estimated by assuming that the motif occurs once per sequence in the positive sequences. To accurately remove a PWM signal from the data, we need to re-estimate p for the estimated mixture component $\mathbf{Z}(\hat{\theta})$. Following Equation 2.1, we have:

$$X - B \approx p(Z(\hat{\theta}) - B) \tag{2.13}$$

(with convergence as $n \to \infty$, if $\hat{\theta}$ is a consistent estimator) since we assume the difference between the observed k-mer counts and the background model results from the PWM. Given our current estimate of $\hat{\theta}$ we can recover p from the above equation. To increase signal, we only use the top 500 k-mers predicted by $Z(\hat{\theta})$ for this computation. Once estimates $\hat{\theta}$ and $\hat{p}$ are determined we assign new values to the number of observed k-mers in the positive and negative sets by setting $X' = X - pZ(\hat{\theta})$ and then all entries are rescaled to sum to 1.

2.3.7 Speeding up the calculation and search

Although the running time of DECOD does not depend on the size of the input dataset, it grows exponentially with k, the length of the motif, since the target function (Equation 2.8) includes a summation over all possible k-mers. Moreover, the search space for the k-mers from which a motif is constructed also grows exponentially with k. In the software implementation, the followings are implemented as an option to speed up the optimization process.

To speed up the calculation of the target function particularly for larger ks, we alternatively first calculate the frequencies of all k-mers in the positive and negative sets, and then the summation in Equation 2.8 is calculated only over those k-mers whose frequency differences are more than 2 (if $k < 10$) or 3 (if $k \geq 10$) standard deviations (sd) away (both sides) from the mean of all k-mers, with the underlying assumption that those k-mers whose frequency differences are small are likely to contribute little to the calculation of the target function.

To speed up the search process, we also limit the initial search space to those k-mers whose frequency differences are more than 1 sd away from the mean. Moreover, we perform two rounds of searches in each iteration. The first round is crude search in which we only use the partial derivatives in Equation 2.9 to estimate the change brought about by adding a k-mer to or removing a k-mer from the motif without doing exact calculation of the target function at all. After a set of m k-mers (m is the motif cardinality) are obtained from the crude search that leads to a motif θ with the maximum target function score at this stage, we expand this set by including all other k-mers that are similar to θ (i.e. in the "neighborhood" of θ). Specifically, the probability of each k-mer given θ is calculated and all k-mers whose probabilities are higher than $0.5^{k/2} \cdot 0.1^{k/2}$ are added to this set. Then a second round of refined search is performed according the optimization process for exact calculations as described in Section 2.3.4, using this set as the new search space. The final motif found by this second round of search is reported.

2.3.8 Method: Using more complicated motif models

In the above we have been using a simple PWM model for representing the motif, which assumes independence between the positions within a motif. Theoretically any motif model that can be constructed from a set of k-mers and can output the probability of observing a k-mer given the motif model can be used. To further test if more complicated models that capture more dependence between positions in a motif lead to better performance, we implemented a first-order Markov motif model which is denoted as DECOD-Markov below. The parameters for the Markov model is learned in the same way as those for the PWM model on the k-mers in the current site set. For each k-mer $a = a_1 \cdots a_k$ and with the Markov model θ,

$$\Pr(a|\theta) = \theta_{1,a_1} \prod_{i=2}^{k} \theta_{i,a_i,a_{i-1}} \tag{2.14}$$

in which $\theta_{i,a_i,a_{i-1}}$ is the probability of observing the nucleotide a_i at position i given the previous nucleotide at position $i-1$ is a_{i-1}, and θ_{1,a_1} is the starting probability of observing a_1 at the first position. Note that with a more complicated motif model like this, using the approximation (Equation 2.9) in the add step of the hill climbing is no longer possible, so we do full calculation in the add step to get the target function score resulting from adding a k-mer to the site set, and the k-mer giving the highest target function score is added. To make up for the additional running time from this, we use the simple mixture component in the target function (Equation 2.7) instead of the convolved mixture component (Equation 2.8).

2.4 Method: Evaluation and comparison on simulated and real data

2.4.1 Motif discovery on simulated data

For each simulated study, 100 simulated datasets were generated and results were averaged. In each dataset, two groups of positive and negative sequences of length 400bp each were first generated using a multinomial background distribution with equal probabilities for A, C, G and T respectively. Then, in the positive set, palindrome motif(s) of the specified width were planted at randomly chosen positions. The information content of a column (column IC) in the PWM is defined as

$$IC = \sum_{i \in \Sigma} f_i \log_2 \frac{f_i}{b_i} \tag{2.15}$$

in which $\Sigma = \{A, C, G, T\}$, f_i is the base frequency of nucleotide i in that column of the motif, and b_i is the base frequency of nucleotide i in the background which is always 0.25 in our case. We compared our method with other popular software specifically designed for discriminative motif finding including: ALSE (v1.07, [93]), Seeder (v0.01, [49]), DME (v2 beta 2008.08.30, [172]), DEME (v1.0, [143]), DIPS (v1.1, [169]) and CMF ([109]), in terms of the accuracy of the recovered motif and the running time needed. For all cases, the accuracy was measured by the average Kullback-Leibler (K-L) divergence per column (AKLD) between the recovered motif and the known planted motif defined as

$$d = \frac{1}{k} \sum_{i=1}^{k} \sum_{j \in \Sigma} (M_{ij} - M'_{ij}) \log_2 (\frac{M_{ij}}{M'_{ij}}) \tag{2.16}$$

in which k is the motif length, $\Sigma = \{A, C, G, T\}$, and M_{ij} and M'_{ij} are the corresponding positions in the two motifs being compared [172]. One position shifting was allowed in calculating AKLD, i.e. when comparing two motifs A and B of length k, three AKLDs were calculated over (i) the full length, (ii) the first $k-1$ columns of A with the last $k-1$ columns of B and (iii) the last $k-1$ columns of A with the first $k-1$ columns of B. The lowest among the three was reported. All methods were run on both strands of the input sequences. For DECOD, on each dataset, both exact and speedup calculations (referred to as “DECOD-exact” and “DECOD-speedup”) were run for 50 iterations (the default value) respectively, and the motif with the best discriminative score was reported. The motif cardinality was set to 20 and the probability of motif occurrence was set to once per positive sequence (the default values) for all analyses, unless otherwise noted. For DME, the option ‘-n 200’ was used to allow the program to return many motifs as suggested in its documentation, and for finding bimodal motifs the option ‘-i 0.5’ was used to allow the program to search for column types with information content as low as 0.5. For ALSE, the option ‘-b’ was used to specify the number of motifs to be 1 or 2 accordingly for each comparison. For CMF, the option ‘-d 1’ was used to set the motif enrichment to be only in the positive sequences and the option ‘-w 6 -l 6 -u 6’ was used to set the length of the motif to be 6. For Seeder, the seed width was set to be 6. For DIPS, we compared running it for both 5 iterations (by default) and 20 iterations (by using ‘-niter 20’) (referred to as “DIPS-5iters” and “DIPS-20iters”). Default values were used for the other parameters for all methods. Running times were measured on a computer cluster with 2x Intel Xeon E5620 CPUs at 2.40Ghz and 24GB RAM.

Detailed descriptions about the motifs planted in generating each simulated dataset as discussed in Results is given below.

Single unimodal motifs

One motif with a dominating nucleotide at each position was planted. To more closely mimic real cases, noise was added to each position of the motif so that the information contents (IC) of each column of the motif ranges from 2 bits (corresponding to a completely deterministic motif) to 0.64 bits (corresponding to a probability of 0.70 for the dominating nucleotide and 0.10 for each of the other three nucleotides).

Single bimodal motifs

In this case, the IC for the unimodal positions in the planted motif was 1.15, and the IC for the bimodal positions was 0.53 (the dominating two nucleotides had a probability of 0.45 each and the other two nucleotides had a probability of 0.05 each). We generated 1,000 positive and negative sequences respectively, and one instance of the motif was planted in each positive sequence.

Two motifs

Two different motifs having from 0 to 6 bimodal positions (column IC 0.53) and the rest positions being unimodal (column IC 1.15) were planted in each of the positive sequences at different positions. When carrying out the comparisons, each method was set to report the top 2 motifs. The AKLDs of both recovered motifs to the two planted motifs were calculated, and the recovered motif that had the smallest AKLD to either of the planted motifs was reported as Motif 1 (Figure 2.8, upward), and the other was reported as Motif 2 (Figure 2.8, downward).

2.4.2 Motif discovery on the yeast dataset

For this analysis, probe sequences experimentally determined to be bound by each of the 65 yeast TFs tested in a ChIP-chip assay [63] were downloaded from `http://fraenkel.mit.edu/Harbison/release_v24/final_set/Final_Motifs/` and used as the positive dataset for each TF. The numbers of bound sequences for each TF range from 14 to 195 with a median of 56. A consensus motif for each of the 65 TFs was inferred systematically in [63] and they were used as a gold standard to compare against in our analysis. The widths of these motifs range from 6 to 18 with a median of 9. Note that not all bound sequences contained the motif for the corresponding TF (Appendix B). The probes with highest binding p-values for each TF as reported in [63] were collected and used as a negative dataset. The number of probe sequences in the negative set is twice the number of those in the positive set for each TF. Then each method studied was run on both strands of the input sequences to search for one motif of the known width for each dataset.

2.4.3 Motif discovery on eukaryotic benchmark dataset

For this analysis, a benchmark dataset [184] was downloaded which contains binding sites for 52 TFs from yeast, fly, mouse and human as well as negative control sets in which no true TFBSs exist [111]. We did not include the yeast data in this dataset in our study since we already performed the comparison on Harbison's dataset. For each TF, the sequences containing the motif within the original genomic context (the "real" background type) were used as the positive sequences, and twice as many randomly selected sequences from the other TFs in the same species were used as negative sequences. The motif width given as input to each motif finding method was specified to be the minimum width of the true binding sites of that TF, as this should be the most informative part of the motif. All methods were set to search for motifs on both strands of the input sequences. For DECOD, only speedup calculation was used. After a motif was found, it was converted to a log-odd scoring matrix using the background frequencies from the positive sequences, and then used to scan both strands of each positive sequence. To allow for flexibility, all k-mers with a score higher than 70% of the maximum possible score for the log-odd scoring matrix were reported to be motif instances [63]. For the other methods, motif instances reported in their output files were used directly. The prediction results for all methods were formatted as required and submitted to the server at `http://bio.cs.washington.edu/assessment/` for evaluation. We focused on two metrics: the nucleotide level sensitivity (nSn) and the nucleotide level positive prediction (nPPV). They are defined as follows:

$$nSn = nTP/(nTP + nFN) \tag{2.17}$$

$$nPPV = nTP/(nTP + nFP) \tag{2.18}$$

in which nTP is the number of true positive predictions, nFN is the number of false negative predictions and nFP is the number of false positive predictions (all at nucleotide level). See [184] for detailed explanations.

2.5 Method: ChIP-chip experiment of the p53 mutant binding

2.5.1 P53 array design

The p53-focused array was designed as previously described [162]. The array includes 540 p53-PET sites, 62 additional previously described p53 target regions and 846 randomly chosen promoter regions. Each spot contains PCR product of the designated region with an average length of about 800 bps.

2.5.2 Cell growth and treatments

H1299 tet-off inducible cell lines were created as previously described [28]. The cells were grown in DME-M (Sigma) supplemented with 10% FCS, 2.5ug/ml tetracycline (Teva), 300 ug/ml G418(Mercury). The wild-type p53 expressing cells had 2 ug/ml puromycin in the culture medium and the 6KQ/6KR cells were cultured with 100 ug/ml hygromycin (Roche) in the medium. p53 induction was achieved by omitting tetracycline from the medium for 24 hours followed by three washes with PBS and either incubation of (6KQ) cells for 24 hours with 2.5 ng/ml tetracycline or wild-type and 6KR cells with 5 ng/ml tetracycline. The levels of p53 in these three clones were similar to each other as determined by Western blotting (Section 2.5.3) and were also similar to the amount of p53 in HCT116 cells treated with 375 uM 5-fluorouracil for 6 hours.

2.5.3 Western analysis

The cellular lysates were separated on 10% polyacrylamide gel, with equal protein amounts loaded on the gel for each sample, then transferred to a nitrocellulose membrane and incubated with mouse anti-p53 (DO-I; Santa Cruz), goat anti-β-actin (I-19; Santa Cruz) antibodies and horseradish peroxidase-conjugated secondary antibodies. The signal was visualized via enhanced chemiluminescence reaction and exposure to film (Figure 2.2)

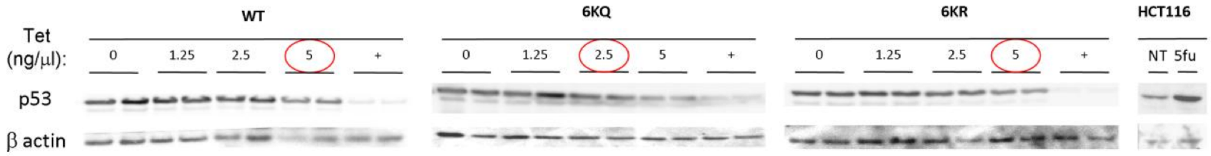


Figure 2.2: H1299 cells with inducible wild type and mutant p53. Western analysis of p53 and beta actin protein levels in WT p53, 6KR p53 and 6KQ p53 H1299 cell clones induced with different concentrations of tetracycline (0-5 ng/ml tet). Cells treated with high amounts of tetracycline (2500ng/ml) to shut off p53 expressions are designated by a plus sign. The final concentration of tet used for induction of p53 in the ChIP analysis is 5 ng/ml for wt and 6KR mutant containing H1299 cells and 2.5 ng/ml for 6KQ mutant containing cells. HCT116 cells without treatment (NT), or treatment with 5-uorouracil (5FU) are shown for comparison of p53 protein levels.

2.5.4 Chromatin immunoprecipitation-on-chip

Chromatin immunoprecipitation (ChIP)-on-chip analysis was performed essentially, as previously described [88], using 10 μg anti-p53 antibody DO-1 (Santa Cruz). Approximately 5×10^7 cells were used. The array was scanned and analyzed with GenePix Pro software, and the fluorescence intensity in both channels was obtained for each spot. As the array is spotted four times, median Cy3 and Cy5 intensities were calculated for each spot. The two channels were normalized according to the median intensity of the random human promoter spots, and the Cy5/Cy3 ratio of each spot was calculated. The experiment was performed in duplicate, and the average binding ratio for each spot was calculated. The significance of the enrichment observed in each spot was determined by calculating the deviation of each ratio from the mean of the random promoters control spots (Z score). Only $\sim 1\%$ of the random promoters obtained Z of > 2.5; thus, this cutoff is equivalent to an FDR of 0.01. For gene-specific validation (data not shown), the ChIP assay was performed as described above and the nonamplified immunoprecipitation and input fractions were subjected to 36 cycles of semiquantitative PCR.

For each comparison, sequences identified to be bound by both factors are only put into the negative set. In the WTP53-6KR comparison, there are 81 sequences in the positive set and 255 in the negative set. In the 6KR-6KQ comparison, there are 110 sequences in the positive set and 158 in the negative set. In the 6KQ-control comparison, 147 sequences in the positive set and 36 in the negative set. For motif finding, both strands of the repeat-masked sequences were searched.

2.6 Result: Discriminative motif finding on simulated data

We first tested the performance of DECOD by comparing it to several other discriminative motif finding methods including DME (v2 beta 2008.08.30, [172]), DIPS (v1.1, [169]), ALSE (v1.07, [93]), DEME (v1.0, [143]), Seeder (v0.01, [49]) and CMF [109] using simulated data. For each simulated study, 100 datasets were generated and results were averaged. In each dataset, two groups of positive and negative sequences of length 400bp each were first generated with equal probabilities for A, C, G and T, respectively. Then, in the positive set, palindrome motif(s) of various strength [as represented by the information content of each column (column IC), see Section 2.4.1] were planted at randomly chosen positions. For all cases, the accuracy was measured by the average Kullback-Leibler (K-L) divergence per column (AKLD) between the recovered motif and the known planted motif [172] (see Section 2.4.1). The lower the AKLD, the closer the recovered motif is to the planted motif. In addition, for DECOD, both the exact and speedup calculations were compared (referred to as 'DECOD-exact' and 'DECOD-speedup' hereafter, see Section 2.3). For DIPS, we considered running for 5 iterations and 20 iterations (referred to as 'DIPS-5iters' and 'DIPS-20iters' hereafter).

2.6.1 Single unimodal motif, small input size

We first planted one palindrome motif of width 6 into each positive sequence. Each position of the motif had one dominating nucleotide (thereby unimodal), with column IC ranging from 2 bits to 0.64 bits (Section 2.4.1). One hundred positive and negative sequences were generated respectively for each dataset and we compared the ability of each method to recover the planted motifs. When the planted motif was strong with a column IC$\geq$1.58, most methods including DECOD-exact, DECOD-speedup, DME, DEME, DIPS-5iters, DIPS-20iters and CMF (to a lesser extent with larger variance) were able to accurately recover the planted motif (average AKLD leq1, Figure 2.3A). However, when the column IC was reduced to 1.15, using CMF, DIPS-5iters and DIPS-20iters led to an AKLD higher than 1, while DECOD-exact, DECOD-speedup, DME and DEME still performed well and were also stable (AKLD$\leq$0.6 with small variance, Figure 2.3A). When the column IC was further reduced to 0.64, the planted motif instances became too noisy with few instances preserving the dominating positions of the motif, and with the small number of sequences available, virtually all methods except ALSE failed (AKLD$\geq$2, Figure 2.3A). However, among all the tested methods, ALSE and Seeder performed poorly when the planted motif was strong. For Seeder, its weak performance for the strong-planted motifs may have been related to the motif length. The seed width for Seeder as input should be shorter than the motif width, but in this case the two were set to be equal since the minimum possible seed width for Seeder was 6. For ALSE, the apparent decreasing AKLD with weaker motif was because ALSE reported matrices in which the distribution at each column is diluted (e.g. [0.5 0.167 0.167 0.167] instead of [1 0 0 0]). In terms of running time, DEME and DIPS required a long time to run ($\sim$15 min for DEME, $\sim$25 min for DIPS-5iters and $\geq$1.5 h for DIPS-20iters, Figure 2.3B). In contrast, DECOD (particularly the speedup version) and DME were the fastest taking $\leq$1 min.

To further mimic real cases in which the motif of interest does not necessarily exist in all positive sequences, we generated simulated datasets in which only some of the 100 positive sequences (percentage denoted as q, varying from 50% to 90%) contained the planted motif (with

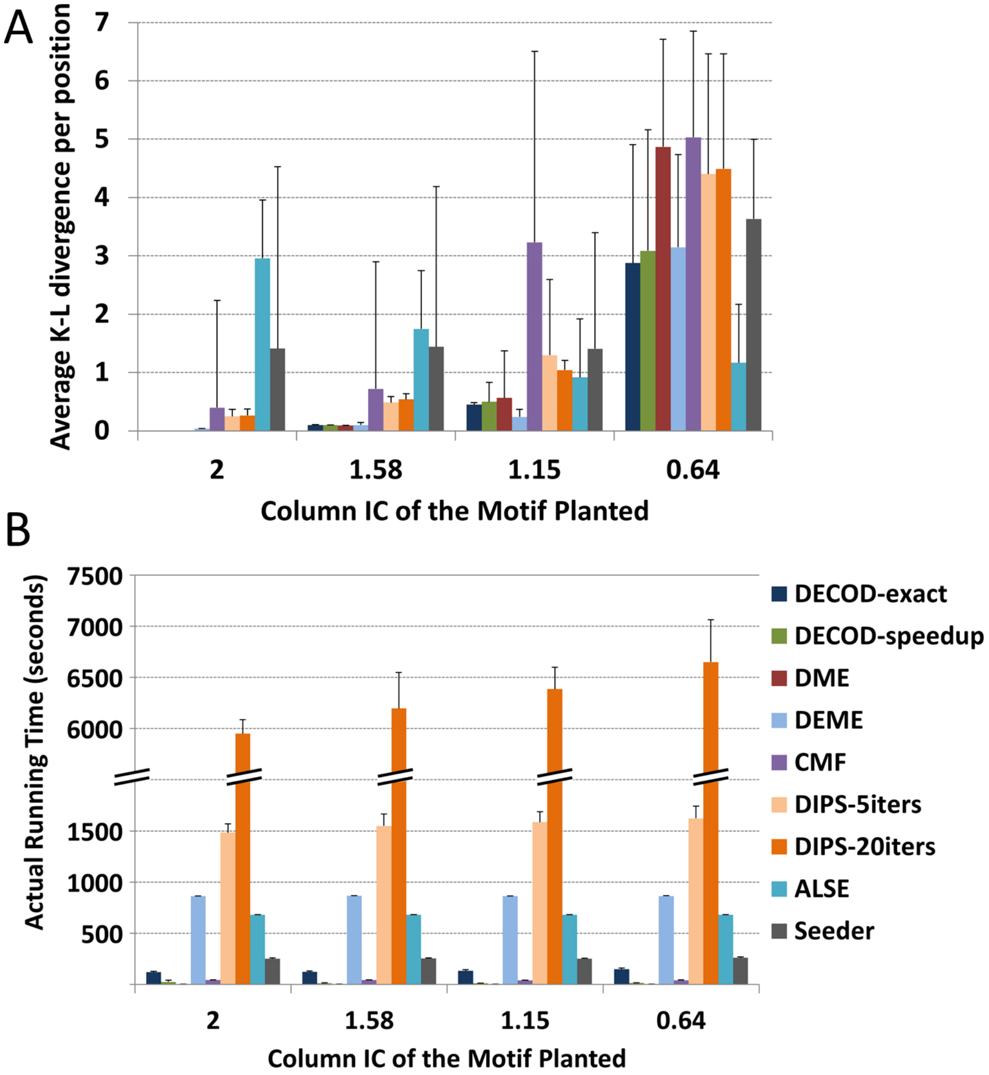


Figure 2.3: Performance comparison on the simulated data planting one motif of width 6 in each of the 100 positive sequences. (A) Average accuracy as measured by AKLD (B) Actual running time. The error bars represent standard deviation based on results from 100 datasets.

an column IC of 1.15). In all the ranges of q tested, DECOD-exact, DECOD-speedup and DEME outperformed the other methods, including DME, in terms of the accuracy of the recovered motif (Figure 2.4). Note that the running time for DEME was more than 15 times longer than DECOD (Figure 2.3B). When q was high ($\geq$0.8), both DECOD-exact and DECOD-speedup had a small variance suggesting that their performance was relatively stable. In contrast, DME had a much larger variance, indicating that it failed on a lot more of the 100 simulated datasets than DECOD (Figure 2.4).

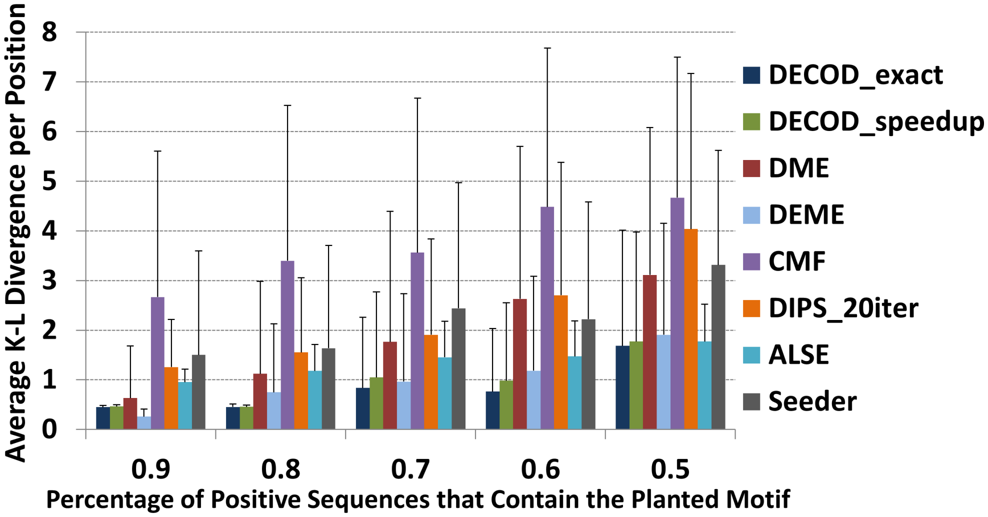


Figure 2.4: Performance comparison of accuracy as measured by AKLD on the simulated dataset in which the motif is only planted in some (x-axis) of the 100 positive sequences.

In order to better evaluate the ability of DECOD to find longer motifs, we further generated simulated datasets in which a palindrome motif of length 8 is planted in some (not all) of the positive sequences (the other settings remain the same as above). We compared the top motif recovered by each method in terms of the AKLD to the known motif and each method's running time. As shown in Figure 2.5A, in all ranges of q tested, the ALKD of the motif that DEME recovered is slightly better than DECOD, which is in turn slightly better than DME. All the above three methods perform much better than the other methods we tested. However, in terms of running time, DEME took about 6 times longer than DECOD (Figure 2.5B). In addition, we did not include DIPS in this comparison due to its excessive running time (≥2hrs for each run).

2.6.2 Single unimodal motif, large input size

To investigate how well each method scales with the size of the input data, we next increased the number of sequences for each dataset to 1,000, and we still planted one motif with varying

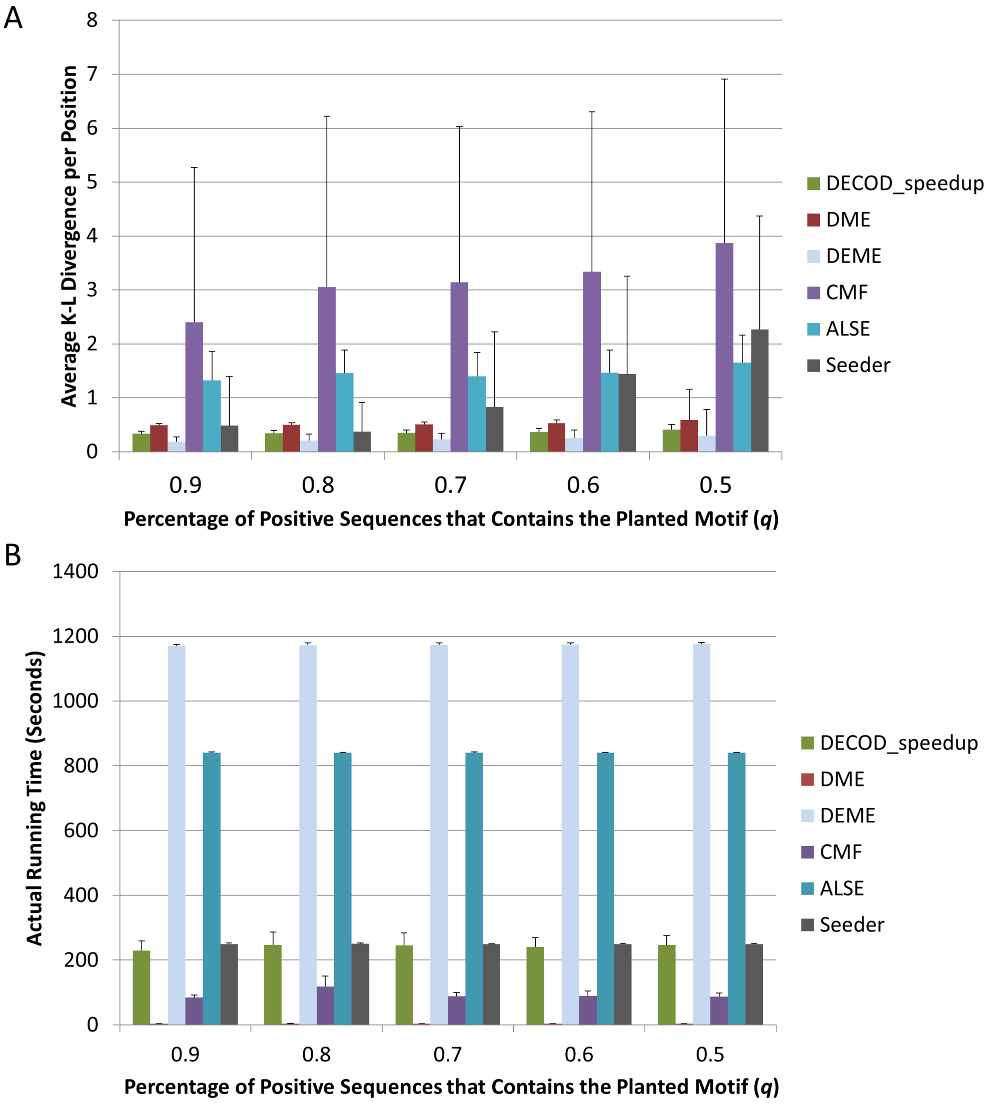


Figure 2.5: Performance comparison on simulated dataset in which a motif of width 8 was planted in some (percentage q, ranging from 50% to 90%) of the 100 positive sequences. (A) Accuracy as measured by AKLD (B) Actual running time (seconds)

column IC in each positive sequence. With this large input dataset size, DIPS failed to run, the running time for DEME and Seeder became prohibitively long (≥6h), and the running time for ALSE increased to more than 1.5 h. We thus excluded them from the analysis and only compared DECOD-exact, DECOD-speedup, DME and CMF. All four methods were able to precisely recover the planted motif and the AKLDs were very similar for all the methods, although the AKLDs increased with lower column IC of the planted motif as expected (Figure 2.6A). In terms of running time, DECOD-speedup and DME were the fastest (≤1 min), followed by DECOD-exact (∼2 min) and CMF (∼6 min, Figure 2.6B).

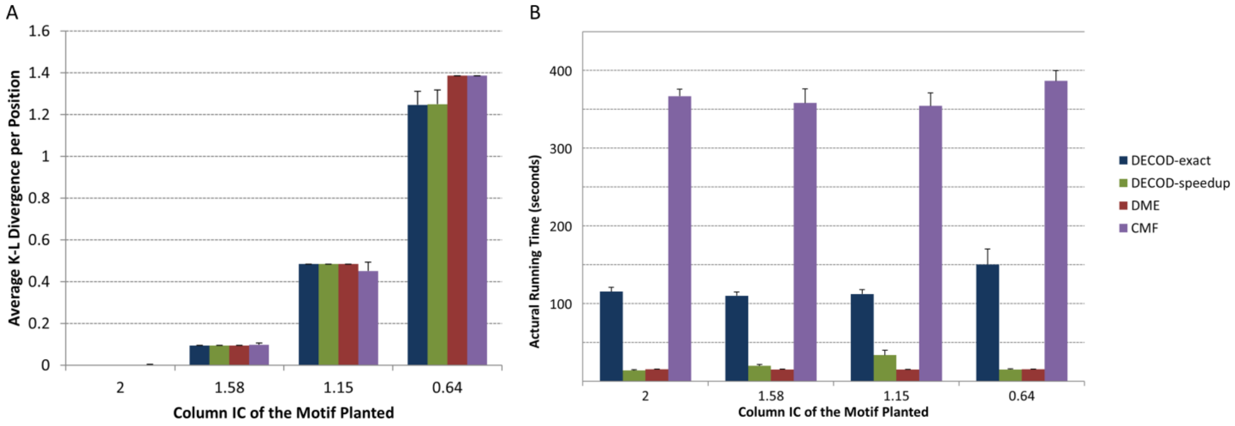


Figure 2.6: Performance comparison on the simulated data planting one motif in each of the 1000 positive sequences. (A) Accuracy as measured by AKLD (B) Actual running time (seconds)

2.6.3 Single bimodal motif

We next tested a more difficult case where some positions (ranging from one to all six) in the planted motif are bimodal (column IC 0.53, see Section 2.4.1). Such cases, in which a motif contains a few weak positions, are very common in practice. As the number of bimodal positions increased, the recovered motifs by all methods tended to diverge further from the planted motif (Figure 2.7). However, the AKLDs of the motifs recovered by both DECOD-exact and DECOD-speedup were in most cases comparable to DME and both were better than CMF.

2.6.4 More than one motif per sequence

In real data, genes are often combinatorially regulated by multiple TFs. To test the ability of our method to recover more than one motif from a dataset and compare with the other methods, we next planted two different motifs in each positive sequence in a simulated dataset containing 1000

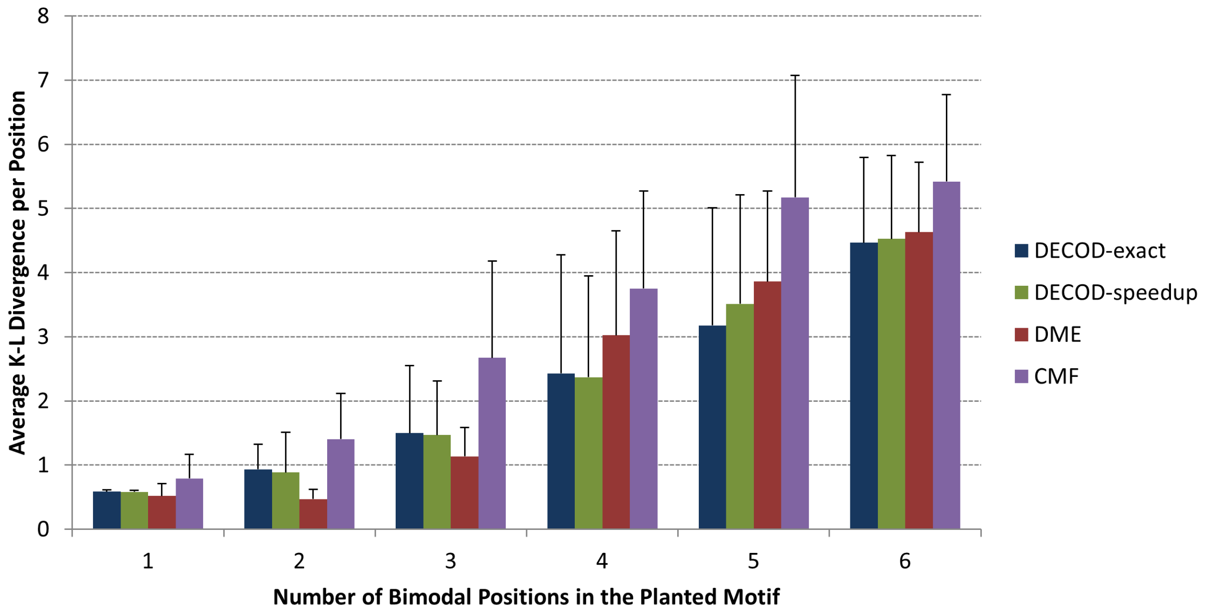


Figure 2.7: Performance comparison on recovering bimodal motifs. A motif containing bimodal positions is planted in each of the 1000 positive sequences

sequences each. The planted motifs had 0-6 bimodal positions (column IC 0.53) and the other positions were unimodal (column IC 1.15) (see Section 2.4.1). Both DECOD-exact and DECOD-speedup were able to correctly recover the two planted motifs (AKLD$\leq$1) and outperformed DME, especially when the number of bimodal positions were $\geq$3 (Figure 2.8 and Section 2.4.1). Interestingly, CMF was able to correctly recover one of the two motifs in most cases and always failed to recover the other (Figure 2.8, downward bars representing the recovered motif with larger AKLD, see Section 2.4.1).

2.6.5 Robustness of DECOD to parameters

To investigate whether DECOD is sensitive to the choice of the motif occurrence probability (p) and cardinality (C) parameters, we ran DECOD on simulated data using a range of different values for these parameters. We also tested the ability of DECOD to predict motifs longer than 6 (k$\geq$6). Similar as before, 1,000 simulated positive and negative sequences, length 400bp each,

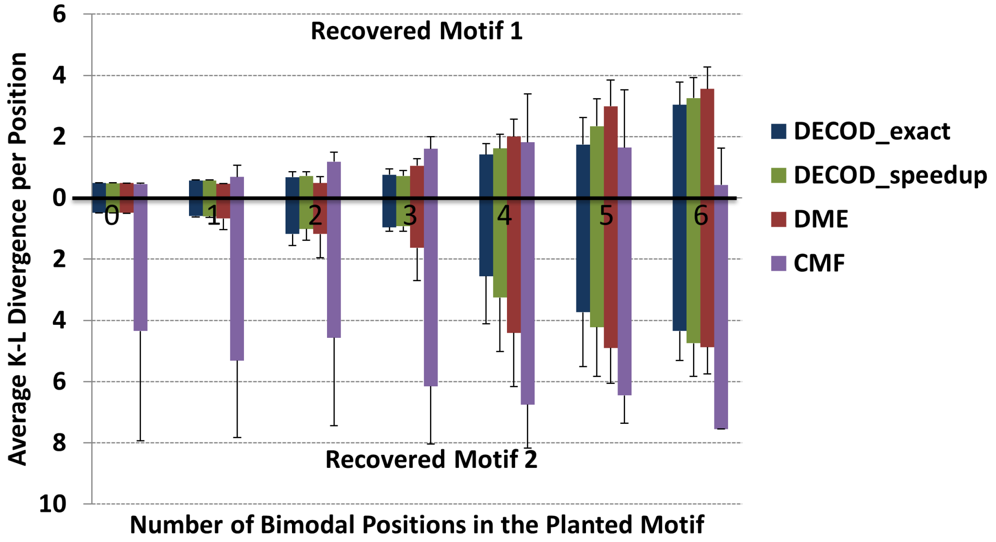
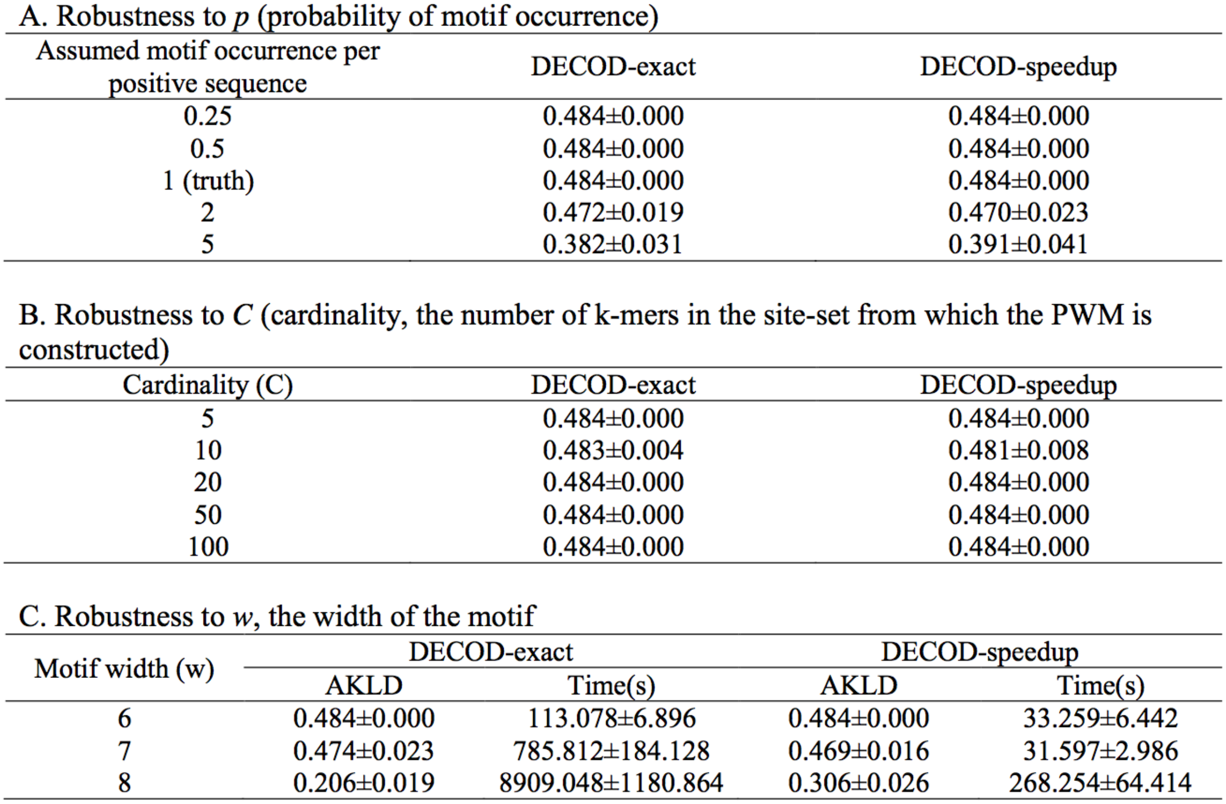


Figure 2.8: Performance comparison of accuracy as measured by AKLD on the simulated dataset in which two motifs were planted in each of the 1000 positive sequences. Each method was set to report 2 motifs. Upward, the AKLD of the recovered motif closer to the two known motifs. Downward, the AKLD of the other recovered motif to the corresponding known motif.

were generated. One motif with column IC 1.15 was planted once in each positive sequence. DECOD was able to successfully recover the planted motif starting with a p as low as 0.25 or as high as 5 times per positive sequences (Figure 2.9A). In reality, some motifs may be more likely to occur more than once in a positive sequence. The insensitivity of DECOD to the value of p suggests that DECOD has the advantage of still being able to correctly recover the motif in such cases. We suggest assuming one occurrence per positive sequence as a starting point. Second, motif cardinality might affect the resolution of the recovered motif. However, for strong motifs as used in our experiments, DECOD works well for C ranging between 5 and 100 (Figure 2.9B). Increasing it to 100 does not affect the result much, though it does increase the run time (not shown) since the search process will necessarily take longer time to converge. On the other hand, with a small C the method is more likely to be stuck in a local optima due to the reduced resolution. Therefore we used $C = 20$ in all further analyses which is also the default choice of Sinha in DIPS [169]. In the command-line version of the program (downloadable from the

Supplementary Website), both of the above parameters (the probability of motif occurrence and motif cardinality) can be user-specified.



A. Robustness to *p* (probability of motif occurrence)

Assumed motif occurrence per positive sequence	DECOD-exact	DECOD-speedup
0.25	0.484±0.000	0.484±0.000
0.5	0.484±0.000	0.484±0.000
1 (truth)	0.484±0.000	0.484±0.000
2	0.472±0.019	0.470±0.023
5	0.382±0.031	0.391±0.041

B. Robustness to *C* (cardinality, the number of k-mers in the site-set from which the PWM is constructed)

Cardinality (C)	DECOD-exact	DECOD-speedup
5	0.484±0.000	0.484±0.000
10	0.483±0.004	0.481±0.008
20	0.484±0.000	0.484±0.000
50	0.484±0.000	0.484±0.000
100	0.484±0.000	0.484±0.000

C. Robustness to *w*, the width of the motif

Motif width (w)	DECOD-exact		DECOD-speedup	
	AKLD	Time(s)	AKLD	Time(s)
6	0.484±0.000	113.078±6.896	0.484±0.000	33.259±6.442
7	0.474±0.023	785.812±184.128	0.469±0.016	31.597±2.986
8	0.206±0.019	8909.048±1180.864	0.306±0.026	268.254±64.414

Figure 2.9: Robustness of DECOD to various parameters. (A) Robustness to the probability of motif occurrence parameter p (B) Robustness to the motif cardinality parameter C (C) Robustness to the motif width parameter k. AKLD, Average K-L divergence per position.

DECOD also works well with longer motifs (Figure 2.9C, see also Section 2.6.1 on simulated data, Section 2.7.1 on yeast data and Section 5.2.1 on ChIP-seq data). Since the exact calculation includes a summation over all k-mers, the running time using exact calculation increases exponentially with k, and therefore when k is too large (e.g. longer than 10), exact calculation is impractical. However, the speedup calculation does not suffer from this since it only makes use of those k-mers that show the most frequency difference between the positive and negative set (Section 2.3.7 and Figure 2.9C), and the accuracy of the speedup calculation is comparable in almost all cases to the results from exact calculations (Figure 2.9C).

2.7 Result: Performance comparison on recovering motifs from biological benchmark datasets

2.7.1 Discriminative motif finding on yeast benchmark dataset

We next applied DECOD to identify transcription factor binding sites (TFBSs) in real biological datasets. For this purpose, we first used a benchmark dataset in *Saccharomyces cerevisiae* [63] and compared DECOD's results with the other methods. For each of the 65 TFs with high-confidence known motifs determined as reported in [63], the probe sequences bound by the TF were used as the positive set (Section 2.4.2). Note that not all bound sequences contained the motif for the corresponding TF. Negative datasets were constructed for each TF by using the probes most unlikely to be bound (Section 2.4.2). We then run each method to search for one motif of the known width for each dataset, and we matched the motifs discovered against a database containing all the motifs for those TFs reported in [63] using STAMP [106]. A discovered motif is considered to be correct if the true TF is within the top 5 matches returned by STAMP. We did not include DIPS in our comparison due to its prohibitive running time. For our method, we only used the speedup version since many motifs are longer than 8.

Out of all the 65 motifs, DECOD was able to recover 28, compared to 31 for DME and 34 for DEME (none of the other methods correctly recovered more than 34 motifs, Table 2.1 and Appendix B). However, the motifs for these 65 TFs are not equally reliable. An enrichment score for each motif was calculated in [63] to measure the relative enrichment of the motif in the bound probes compared with all intergenic sequences in yeast. Motifs with a higher enrichment score occur more densely in the bound probes and are therefore more reliable. Of the 21 motifs with an enrichment score ≥25, DECOD was able to recover 15, similar to the number recovered by DEME (also 15) and higher than the number recovered by DME (13) (Table 1). It should be noted that many of the motifs correctly recovered by DECOD are longer than 10 (B), and that the running time for DECOD is always much faster than DEME. Therefore, DECOD performs well in recovering yeast motifs from this dataset especially for highly reliable motifs.

Table 2.1: Comparison of discriminative motif finding methods on the yeast dataset

TF	DECOD	DME	DEME	CMF	Seeder	ALSE	Width	Enrichment
ABF1	+ [a]	+	+	+	+		13	99
CBF1	+	+	+	+		+	7	99
FHL1	+	+	+	+	+		10	99
RAP1	+	+	+	+			10	79.92
REB1	+	+	+	+	+		7	77.93
UME6	+	+	+	+	+		8	72.32
RPN4							9	72.02
GCN4	+	+	+	+	+		7	64.62
YAP7							8	62.65
MCM1		+	+				11	55.28
NRG1				+			7	45.42
MBP1	+	+	+	+	+		7	40
SKN7						+	9	38.79
CIN5	+		+	+			8	38.36
SUM1	+	+	+		+		10	36.47
SWI6	+	+	+	+	+		7	33.62
HSF1	+						13	32.96
SWI4	+	+	+	+	+		7	31.96
TYE7	+	+	+	+	+	+	8	30.56
SFP1							9	26.64
FKH2	+		+	+	+		7	26.62
Total (Top)[b]	15	13	15	14	11	3		
Total (All)[c]	28	31	34	24	17	9		

[a] +: Correctly recover the known motif.

[b] Total (Top) : The total number of the top 21 motifs with enrichment score ≥25 correctly recovered by each method.

[c] Total (All): The total number of all motifs correctly recovered by each method (see Appendix B for details).

We also compared two DECOD variants on this yeast dataset: (1) DECOD-simple, in which we only used the simple mixture component in the target function(i.e. Equation 2.7) instead of the convolved mixture component, in order to see how much improvement the deconvolution brought about; and (2) DECOD-altseed, in which instead of seeding DECOD randomly and performing several restarts with different seeds, we seeded DECOD from k-mers sampled from the known motif for each TF directly. This allows us to identify cases where the original DECOD failed because it did not come up with the best seeds. While the original DECOD recovered 28 of the 65 yeast TFs, DECOD-simple was able to recover 26, and DECOD-altseed recovered 34 (see details in Appendix B). Therefore, deconvolution indeed led to improvements in the performance of DECOD at least for some cases. Moreover, DECOD-altseed recovered several more TFs than the original DECOD. The reason that the original DECOD failed on these cases may be either that an insufficient number of restarts were performed so that the best seeds were not reached, or that the best seeds were completely excluded from the search space due to the limitation of the search space to only those k-mers with extreme frequency differences between the two sets of sequences in the speedup process (Section 2.3.7). For the former case, more random restarts can be performed as a remedy. For the latter case, the limitation on the search space can be relaxed or completely removed in order to allow more k-mers to be considered in the seed.

2.7.2 Discriminative motif finding on eukaryotic benchmark dataset

To further examine the ability of DECOD to discover motifs in more complex organisms, we tested its performance and compared with the other methods using another benchmark dataset [184]. This dataset contains the binding sites for 52 TFs from yeast, fly, mouse and human as well as negative control sets in which no true TFBSs exist. This is a challenging dataset due to the small number of sequences for each TF. Since we already performed a comprehensive comparison on yeast, we only used the 46 TFs from the other three species in this comparison. For consistency with previous analysis of this data, and because unlike the previous data we used in this case the input data contain motif occurrence information, we evaluate the results for each method in terms of the metrics used by [184], particularly the sensitivity(nSn) and positive prediction values (nPPV), at the nucleotide level (Section 2.4.3).

Although the performance for all methods was not great for this subset of the data, at the nucleotide level, DECOD performs better than most of the other methods including DEME, DME and CMF (Figure 2.10). Moreover, the sensitivity (nSn) of DECOD was slightly higher than DEME, DME and Seeder (Figure 2.10). Therefore, DECOD is also competitive with the other methods on recovering TFBS in higher eukaryotes from this dataset. Interestingly, although ALSE performed poorly on the simulated and yeast dataset, it outperforms the other methods on this dataset. Notice that here the nSn and nPPV values we obtained for all methods were generally lower than the results reported in [184] for general (not discriminative) motif finders. This is because in [184] the dataset was sent to the authors of each motif finder software, and many authors performed additional filtering steps, many including inspection by eyes, to improve their predictions. In our scenario we did not attempt to do these because our purpose was only to perform a fair comparison of the methods being evaluated using a consistent standard. We expect that after performing careful post-processing of the motifs reported by each software, the results can be further improved.

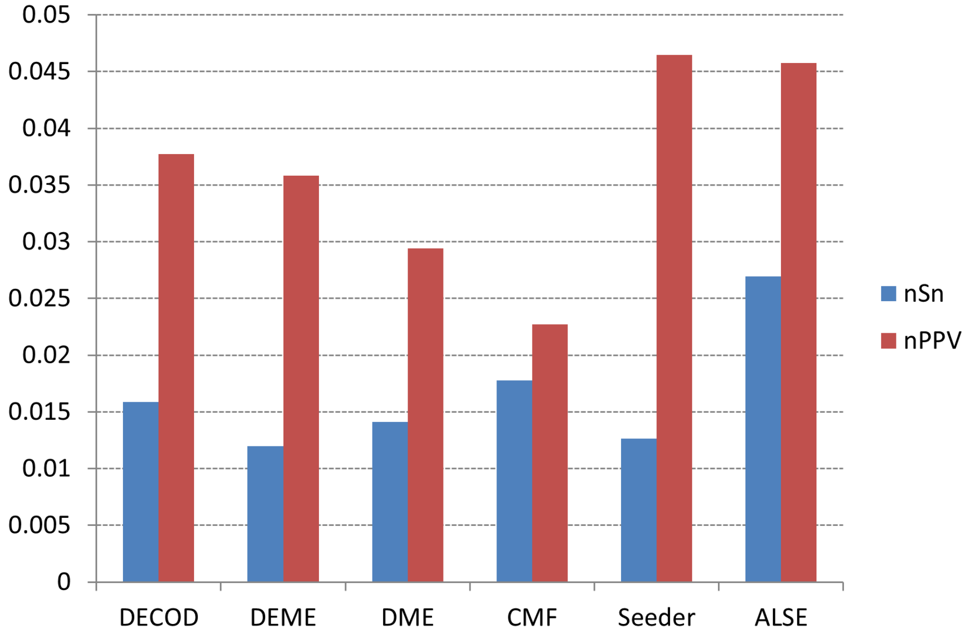


Figure 2.10: Sensitivity (nSn) and positive prediction value (nPPV) at the nucleotide level for each method on the fly, mouse and human TFBS benchmark dataset [184]

2.8 Result: Discriminative motif discovery from p53 mutant binding targets

We next looked at the tumor suppressor p53 in human, a TF which plays a major role in cancer by binding numerous targets [189]. P53 is regulated by many posttranslational modifications, primarily at the amino and carboxyl terminal regions [147]. In particular multiple lysines within the C-terminal domain (CTD) have been reported to undergo numerous modifications including acetylation, methylation, ubiquitination, and SUMOylation [84]. The functions of the acetylation of these lysines have remained elusive. To study the role of the acetylation of these CTD lysines in p53 binding, we performed ChIP-on-chip experiments comparing three H1299 cell lines expressing p53 variants expressed from a tetracycline-regulatable promoter (tet-off) in which the levels of p53 protein can be regulated by varying the amount of tetracycline in the culture medium (Section 2.5). The levels of p53 were calibrated so that equivalent amounts of p53 were expressed in each of the following three cell lines containing: (i) wild-type p53 (WT p53); (ii) mutant p53 in which the six lysine residues in the C-terminus were mutated to arginine (6KR p53), which conserve charge but disallow any lysine modification; and (iii) mutant p53 in which the same six lysines were mutated to glutamine (6KQ p53) which is thought in some cases to mimic acetylation [77] (Section 2.5). To determine their relative affinity for p53 target sites, we used a p53 custom array containing promoters for 600 of p53's targets binding sites.

We found that WT p53 bound to 330 targets, 6KR p53 bound to 255 targets and 6KQ p53 bound to only 150 targets. Interestingly, the 6KQ targets were included in the 6KR targets, which in turn were included in the WT targets. Thus, each of the p53 forms bound a smaller subset of related targets (Figure 2.11A). Since all genes on the array contain a strong p53 binding motif, motif discovery on one target set would lead to the same motif. Thus we used DECOD to search for discriminative motifs that are enriched in one set of these targets versus another. The bound sequences identified in the ChIP-chip experiment in each pairwise comparison of the wild-type p53 and two mutant p53 (6KR and 6KQ) were repeat masked and then used for this analysis. Since the experiment is not strand-specific, both strands were included in the input sequences. We used STAMP [106] to match the motifs we recover with known transcription factor binding sites in the TRANSFAC 11.3 database [111].

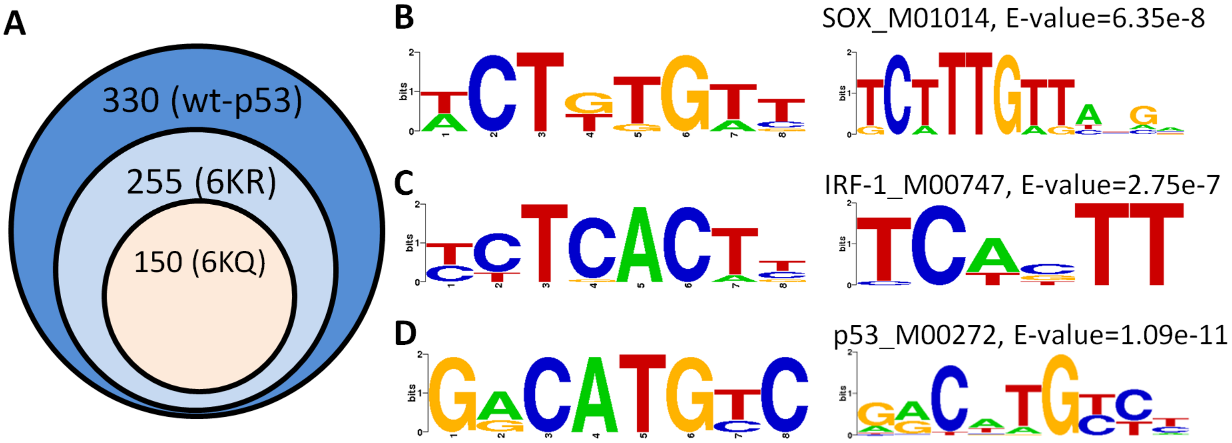

A
330 (wt-p53)
255 (6KR)
150 (6KQ)
B
SOX_M01014, E-value=6.35e-8
C
IRF-1_M00747, E-value=2.75e-7
D
p53_M00272, E-value=1.09e-11
bits

Figure 2.11: Results on the p53 dataset. (A) Number of targets and inclusion patterns for the three p53 forms we tested. (B-D) Discriminative motifs identified by DECOD for the p53 binding datasets. Left: Motifs found by DECOD. Right: matched motifs in TRANSFAC using STAMP [106] (E-values provided by STAMP). (B) The SOX4 motif found in the comparison of the WT p53 targets against the 6KR p53 targets. (C) The IRF-1 motif found in the comparison of the 6KR p53 targets against 6KQ p53 targets. (D) The p53 motif found in the comparison of the 6KQ p53 targets against the control sequences.

DECOD identified several such discriminative motifs in pairwise comparisons between these sets (Figure 2.11B-D). The motifs identified provide new insights regarding co-factors of p53 and the post-translational modification that it undergoes. For example, when comparing targets of WT p53 that are not targets of 6KR p53 to targets of 6KR p53, DECOD identified a motif closely matching the PWM for Sox4 (Figure 2.11B, E-value = 6.35e-8). Sox4 participates in a wide range of cellular processes particularly in cancer [146], and recently it was reported to physically interact with p53 and regulate p53 stability at the protein level [132]. Both the DNA-binding domain (DBD) and the C-terminal domain (CTD) of p53 were shown to be involved in forming the interaction with Sox4 [132]. Since the Sox4 motif was only found to be enriched in comparing the WT p53 targets against 6KR p53 (in which the CTD was mutated) but not in the other comparisons, our result confirms this finding and also suggests that the CTD lysines might be important in maintaining the conformation of the binding site between the p53 and Sox4 proteins. Another example is the motif closely matching the PWM for interferon regulatory factor 1 (IRF-1) when comparing the 6KR p53 against the 6KQ p53 targets (Figure 2.11C, E-value = 2.75e-7). IRF-1 acts synergistically with p53 at the p21 promoter and is coordinately upregulated with p53 during DNA damage response [131]. On the p21 promoter IRF-1 and p53 interact through the p300 acetyl transferase, and this interaction is important for the acetylation of p53 [41]. If p300 is indeed necessary for IRF-1 - p53 interaction, we expect it to be lost after p53 is fully acetylated. Indeed, we found that IRF-1 binding sites are depleted from promoters of the acetylation mimicking mutation (6KQ) raising the possibility that p53 needs the interaction with the IRF-1 protein to control a subset of its targets. Finally, in comparing the 6KQ targets against a control set, DECOD was able to recover the motif corresponding to the PWM for p53 (Figure 2.11D, E-value = 1.09e-11). Note that the p53 motif was not found in either of the previous comparisons due to the discriminative nature of the method, which is what we desired since all the three sets of targets contains the motif.

To see if the other methods can recover these motifs, we also run DME and CMF on this dataset. Each method was set to search for 10 motifs of width 8 in each comparison, and all methods were run on both strands of the repeat-masked input sequences (same as used for DECOD). The results were again compared to known motifs in TRANSFAC [111] using STAMP [106] in the same way we did for DECOD. We did not test DEME and DIPS due to their excessive running time and the fact that DEME is able to find only one discriminative motif. Also ALSE and Seeder were not included since they could not work properly with repeat-masked sequences as in our input dataset. Both DME and CMF were able to find the motif matching

IRF-1 in the 6KR-6KQ comparison (for DME, E-value = 2.59e-6; for CMF, E-value = 1.46e-7; Figure 2.12). However, DME was not able to recover the known p53 motif from the 6KQ-control comparison. Although CMF was able to find a motif similar to that for p53 in the 6KQ-control comparison, the match was very weak (E-value = 1.30e-4, Figure 2.12) compared with the one recovered by DECOD (E-value = 1.09e-11, Figure 2.11D). Neither methods were able to find the motif matching Sox4 that DECOD identified from the WTP53-6KR comparison. The full list of all motifs identified by each method and their matches to known motifs using STAMP is available on the Supplementary Website at http://www.sb.cs.cmu.edu/DECOD/.

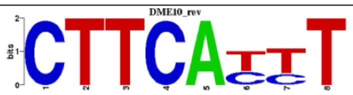
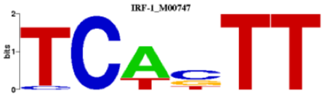
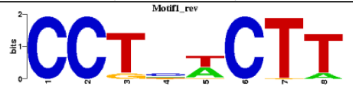
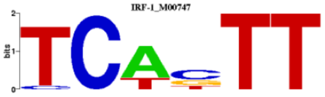
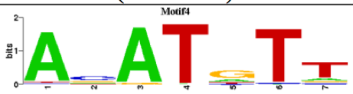
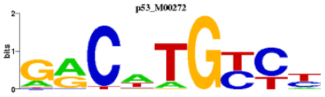
Method	Comparison	Motif recovered by the method	Known motifs in TRANSFAC	Match E-value
DME	6KR-6KQ	 (10th motif)	 IRF1_M00747	2.59e-6
CMF	6KR-6KQ	 (1st motif)	 IRF1_M00747	1.46e-7
CMF	6KQ- control	 (4th motif)	 P53_M00272	1.30e-4

Figure 2.12: Motifs recovered by DECOD and CMF on the P53 dataset (see Supplementary Website for full details)

2.9 Result: Using DECOD to find motifs from ChIP-seq dataset

The running time of DECOD does not depend on the size of the input sequences, therefore DECOD is particularly suited for motif finding from data generated by large scale sequencing efforts such as ChIP-seq experiments. Here we demonstrate the ability of DECOD to recover known motifs for transcription factors from several published ChIP-seq dataset from the EN-

CODE project [44]. Genomic sequences (peaks) determined to be bound in ChIP-seq studies by the following five TFs that have known motifs in JASPAR [22] were downloaded and used as the positive set for each TF: c-Jun (K562 cells), c-Myc (K562 cells), Max (K562 cells), Egr-1 (K562 cells) and NFkB (GM12878 cells). Each set contains tens of thousands of sequences, and the lengths of the peak regions are typically a few hundreds (Figure 2.13). For negative sequences, we used both the upstream and downstream sequences flanking the peak regions, and the length of each flanking sequence was chosen to be the same as the corresponding peak region. Thus the negative set contains twice as many sequences as the corresponding positive set for each TF. We used DECOD to search for motifs of about the width of the known motif (Figure 2.13) on both strands of the input sequences. As shown in Figure 2.13, DECOD was able to correctly recover all 5 motifs even for relatively longer and more complex motifs like CTCF. For 4 of the 5 motifs (c-Jun, Max, CTCF and NFkB), the correct motif that DEOCD recovered was also the first one reported. For c-Myc, the correct motif was the second motif that DECOD recovered. Therefore, DECOD works well on finding motifs from ChIP-seq datasets in which the number of input sequences can be too big for other motif finding software to handle.

In higher organisms like human, it has been suggested that there can be dependencies between positions in a motif that cannot be represented with a PWM model [6, 105]. Here we evaluated using a more complicated first-order Markov motif model that is able to capture dependencies between adjacent positions in DECOD (DECOD-Markov, see Section 2.3.8 for details). In order to evaluate the performance of DECOD-Markov and compare with the original DECOD (DECOD-PWM), we run both on this human TF ChIP-seq dataset. For each of the 5 TFs, the positive and negative sequences were randomly split into two halves. Each method was run on one half to search for a discriminative motif, and the resulting motif models were used to scan sequences in the other half using log likelihood ratio scores. Since both positive and negative sequences are present, we consider this as a classification problem, and we use the area under the ROC curve (AUC) as the criteria to evaluate how well each method works. For

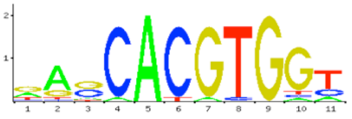
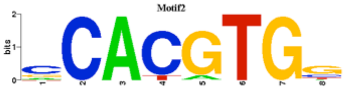
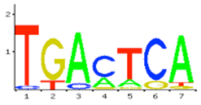
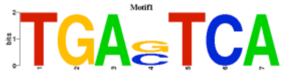
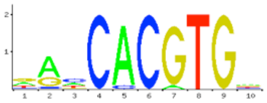
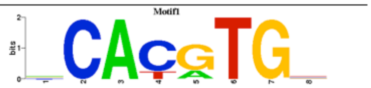
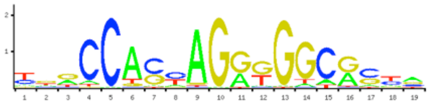
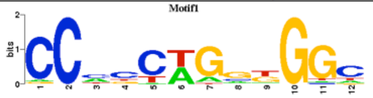
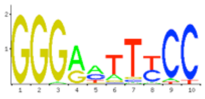
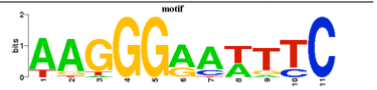
TF	#Seqs[1]	Median length	Known motif (JASPAR)	Recovered motif (DECOD)
c-Myc (K562)	15479	599	 MA0059.1	 (Search for width 8, 2nd, E-value = 4.7e-11)
c-Jun (K562)	26920	384	 MA0099.2	 (Searching for width 7, 1st, E-value = 9.1e-10)
Max (K562)	10480	395	 MA0058.1	 (Searching for width 8, 1st, E-value = 1.96e-5)
CTCF (K562)	64387	151	 MA0139.1	 (Searching for width 12, 1st, E-value = 9.59e-14)
NFkB (GM12878)	38559	493	 MA0061.1	 (Searching for width 11, 1st, E-value = 1.24e-11)

Figure 2.13: Using DECOD to find motifs from ChIP-seq datasets. [1]The ChIP-seq peak regions were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/

DECOD-Markov, in order to avoid over-fitting, we considered both using a motif cardinality of 20 and 100. In Figure 2.14 we show the AUC and actual running times for each method on c-Jun, c-Myc and Max that have widths 7-8. As can be seen, in all three cases the AUC achieved by the Markov motif model for both motif cardinalities of 20 and 100 is very close to that achieved by DECOD-PWM. However, the running time for DECOD-Markov (several hours) is much longer than that for DECOD-PWM (several minutes). In addition, in Figure 2.15 we show the AUC of DECOD-PWM on the two longer TFs, CTCF and NFkB. For these longer motifs DECOD-PWM takes 1.2 hours (NFkB, 11bps) and 4.3 hours (CTCF, 12bp) respectively, but DECOD-Markov did not finish within 7 days for either. Therefore, overall at least for this specific dataset, the Markov motif model does not lead to further improvement for DECOD in recovering the known motifs.

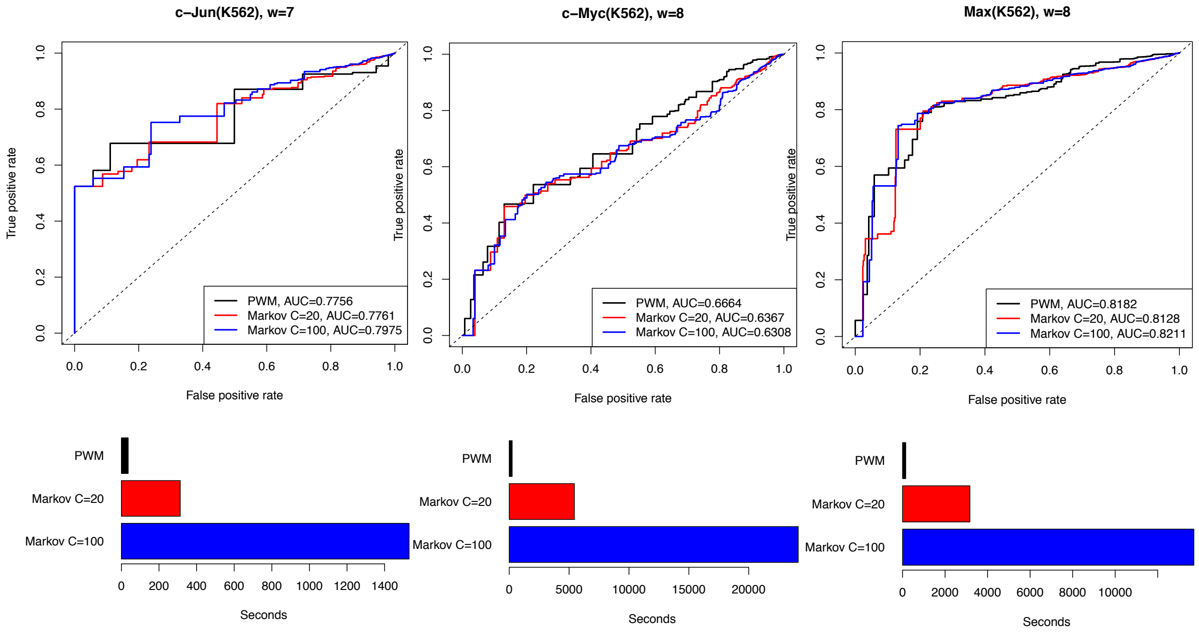


Figure 2.14: Comparison of ROCs and running times for DECOD-Markov and DECOD-PWM on c-Jun, c-Myc and Max. Shown on the top are the ROCs and AUCs for each method, and on the bottom are the actual running times for each method.

2.10 Discussion

We presented DECOD, a novel method for discriminative motif finding in DNA sequences. DECOD uses a deconvolution method which allows it to have a run time independent of the input data size while still taking into account context information.

While DECOD's run time is independent of the input data size, calculating the exact target function (DECOD-exact) increases exponentially with the motif length k. We presented a solution for speeding up the calculation by only using the most informative k-mers (DECOD-

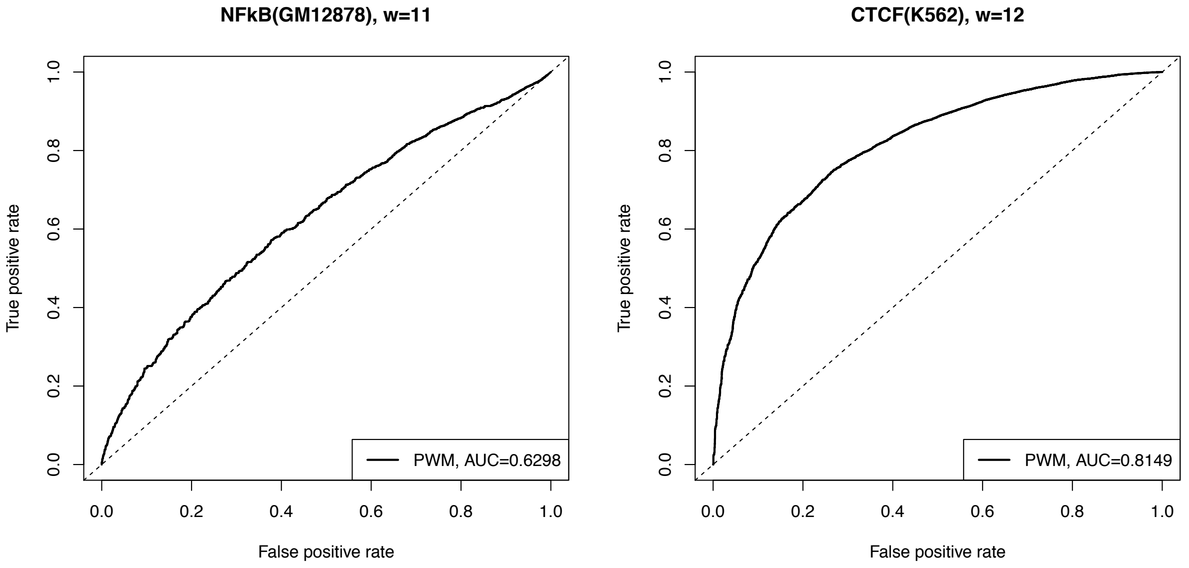


Figure 2.15: ROCs and AUCs for DECOD-PWM on CTCF and NFkB. DECOD-Markov did not finish within 7 days on these longer motifs.

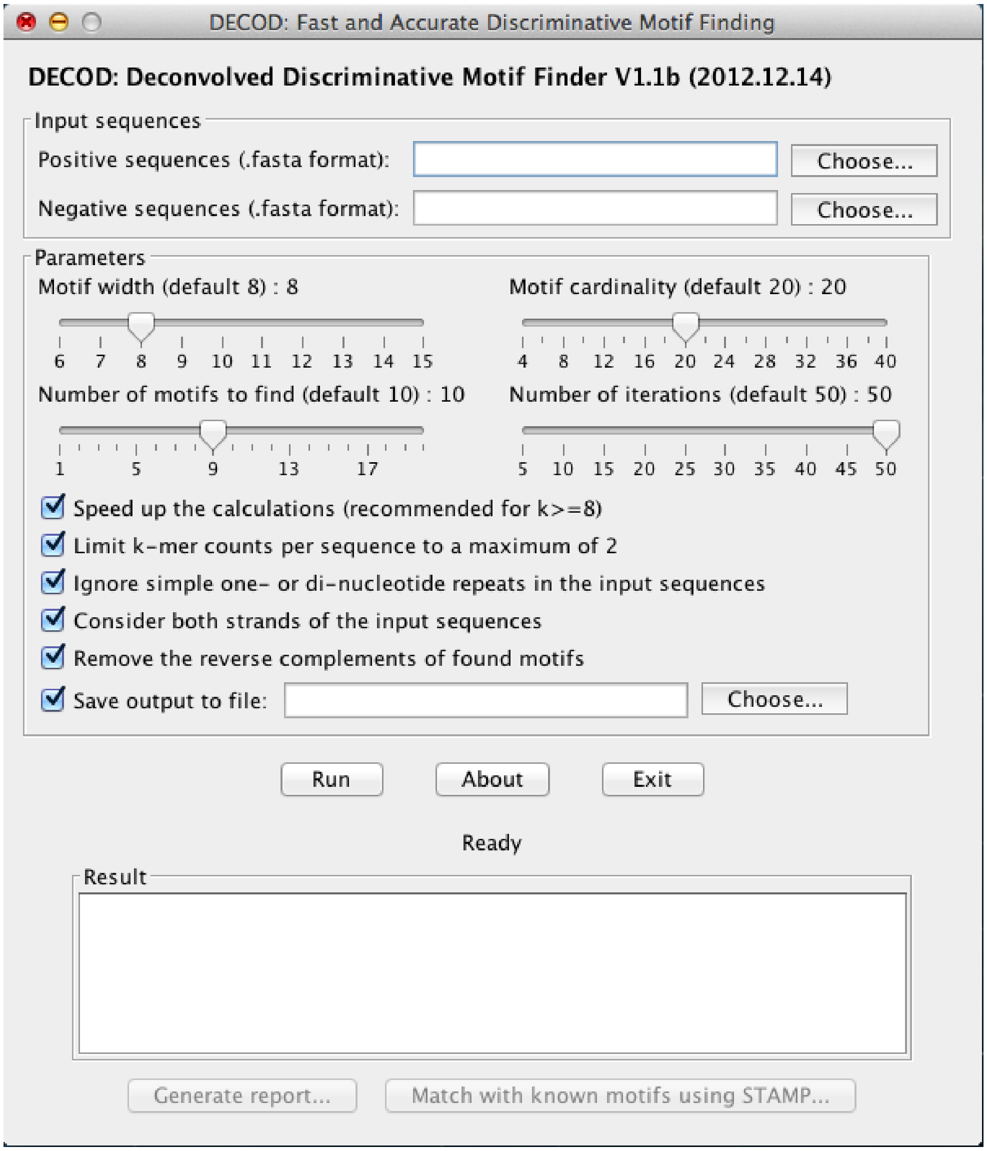
speedup), and showed that it yields motifs that are almost as accurate as those obtained using DECOD-exact while the running time is greatly reduced. As we discuss, DECOD is robust to several input parameters including the choice of the probability of motif occurrence.

When tested on simulated data for which the correct motif is known, DECOD outperforms all other methods when searching for complicated motifs with bimodal position and when looking for combinatorial regulation. It is also much faster than most other methods making it applicable to large sequencing datasets. On real biological benchmark datasets (both yeast and higher eukaryotes), we showed that DECOD was comparable, or better, than other discriminative motif finding methods with the possible exception of DEME for the yeast data. However, as mentioned above, DEME is very slow and so may not be a useful method when studying large datasets. Using DECOD we were also able to identify motifs that are differentially enriched in different p53 mutants which allowed us to identify co-factors of this important TF. Additional experiments are crucial for deciphering the exact interactions between p53 and these other factors, and our bioinformatics analysis using DECOD paves the way for future experiments. We have also tested DECOD using large-scale ChIP-Seq dataset for 5 TFs. For all five DECOD was able to identify the correct motif indicating that it works well on high-throughput datasets as well.

While DECOD was successful in our analysis, it also has limitations. Since DECOD depends on k-mer counts, it does not work well on motifs with large gaps in the middle, since the signals for the k-mers corresponding to the occurrences of such motifs will be more uniform due to the gaps. In future we hope to further extend DECOD to deal with such cases. In addition, when the occurrence of motifs in the positive set is much more sparse than once per positive sequence, DECOD may fail as the default value for motif frequency p is much larger than the actual value. Currently to address this the user is allowed to manually specify the value for this parameter according to their expectations when running DECOD. It would be interesting to develop ways to automatically determine the best value for p. Moreover, we also hope to further improve DECOD by developing ways to automatically determine the length of the motif to be searched for, which can be important when presented with new dataset in which the motif is completely unknown.

2.11 Software availability

DECOD is available for download at `http://www.sb.cs.cmu.edu/DECOD`. DECOD is written in Java and therefore can be run on multiple platforms. It has an easy-to-use graphical user interface (GUI, Figure 2.16), and it can also be run in command line mode for batch processing.



DECOD: Fast and Accurate Discriminative Motif Finding

DECOD: Deconvolved Discriminative Motif Finder V1.1b (2012.12.14)

Input sequences

Positive sequences (.fasta format):

Negative sequences (.fasta format):

Parameters

Motif width (default 8) : 8

Motif cardinality (default 20) : 20

Number of motifs to find (default 10) : 10

Number of iterations (default 50) : 50

☑ Speed up the calculations (recommended for k>=8)

☑ Limit k-mer counts per sequence to a maximum of 2

☑ Ignore simple one- or di-nucleotide repeats in the input sequences

☑ Consider both strands of the input sequences

☑ Remove the reverse complements of found motifs

☑ Save output to file:

Ready

Result

Figure 2.16: Screenshot of DECOD graphical user interface.

Chapter 3

Predicting tissue-specific transcription factor binding

While DECOD is a useful method for *de novo* motif discovery, for many TFs knowledge about their binding is already available. Recently, high-throughput protein binding microarrays were used to measure the binding preferences for hundreds of TFs in an unbiased manner. A key question in these experiments is how to use such data to determine genome-wide binding sites for a specific TF *in vivo*. Such knowledge is important for better understanding transcriptional regulatory networks. In this chapter, we present methods and analyses that provide a solution for this task.

3.1 Introduction

Deciphering transcriptional regulatory network (TRN) requires knowledge about the genome-wide binding sites of transcription factors (TFs) [27, 63]. Chromatin immunoprecipitation(ChIP) followed by microarray (ChIP-chip) [23] or sequencing (ChIP-seq) [134] have been extensively used to interrogate the *in vivo* binding locations of individual transcription factors and coactivators in a wide range of species and tissues [27, 63, 74, 78, 144, 155, 200]. Despite their popularity, such methods require the availability of specific antibody to the TF being studied, and they only study a single cell type, under a specific condition, in one experiment. Thus it is difficult, by using such experiments alone, to obtain a comprehensive understanding of the complicated mammalian TRNs that involve thousands of TFs whose activties change across different tissues and conditions. Computational predictions using other genomic resources to predict tissue-specific transcription factor binding is therefore an important research challenge.

Universal protein binding microarray (PBM) is a high-throughput technique that measures the *in vitro* binding specificity of a sequence-specific transcription factor in an unbiased manner [17, 119], and it has been used to reveal the binding profiles of hundreds of TFs in yeast [207], worm [60] and mouse [6]. Compared with ChIP-based experiments, PBM has the advantage that it does not require a specific antibody to the TF of interest, and is independent of the tissue or condition being studied and so only one experiment is performed for each TF. A number of methods have been proposed for using PBM data to identify TF binding sites (TFBSs). Many of the proposed methods represent TF binding preference extracted from the PBM data by position weight matrices (PWMs) [17, 129, 204]. Such representation, although popularly used due to its simplicity, assumes independence between positions, an assumption which may not hold in many cases [6, 105]. In addition, conflicting results have been reported with regard to whether many TFs have more than one binding preference represented by PWMs [6, 204]. A recent study [129] compared the performances of using PWM representations derived from the PBM data by several methods in predicting *in vivo* binding sites, and concluded that almost all methods performed poorly.

The fact that PBM is an *in vitro* technique is also its biggest weakness. *In vivo* binding is affected not just by motif recognition but also by other tissue-specific conditions that are not observed with PBM experiments including chromatin accessibility, the presence of co-factors, etc [177]. Thus, effective computational methods and additional information are needed when attempting to use the raw fluorescence intensities provided by PBMs to determine tissue- and condition-specific TFBSs. Recently, a number of studies have reported that epigenetic information including certain histone modifications and hypersensitivity to DNase I cleavage correlate with TF binding *in vivo* [73, 95]. Moreover, functional TFBSs tend to be under stronger negative selection, leaving a “phylogenetic footprint” in the genomic sequences. Several methods for predicting *in vivo* TF binding sites have successfully combined such information with PWMs [36, 48, 124, 139]. However, while these approaches led to useful results, relying on PWM-based representation leads to missing real targets as we show in Results. In addition, none of these methods have so far been applied to elucidate the complete set of targets for TFs across a large number of tissues. What is lacking is an integrated model that combines the strength of PBM data with the additional information from epigenetic and/or evolutionary data to predict biologically important TF binding events in multiple tissues.

To address these issues we developed a new method for using PBM data to search for TFBS. Our method, PLAR(*p*ositive *la*sso *r*egularized)-PBM, uses a biophysically-motivated k-mer based model which allows secondary binding profiles and nucleotide dependencies in different position of the TF binding sites. As we show, when predicting *in vivo* TF binding locations determined from ChIP-seq experiments, our method outperforms several other methods suggested for using such data. We then develop an integrative model that combines PBM data with DNase I hypersensitivity data and evolutionary conservation data to predict tissue-specific TFBS *in vivo*. We demonstrate that such an integrative model significantly boosts prediction results compared with using PBM data alone, and show that it improves upon other methods developed to combine such data with PWMs. Finally, we created a resource for tissue-specific TRNs using PBM data for 284 mouse TFs from UniPROBE [126] and DNase I hypersensitivity data for 55 mouse tissue/cell types from the mouse ENCODE project [118]. As we show, many of our predicted tissue specific TFs agree with current knowledge, and global analysis strongly supports our predictions as well. The comprehensive resource of TF binding sites we built thus provides a reference map for understanding complex gene expression patterns.

3.2 Method: PLAR-PBM - Using PBM data to predict TF binding

The binding specificities of TFs are often represented by position weight matrices, which assume that each position of a site contributes independently to the overall binding affinity of the site (independence assumption). The PBM data simultaneously measures binding of a TF to tens of thousands of probes, and can be used to construct a much more detailed and accurate model of TF binding specificities.

3.2.1 A biophysically-motivated model for PBM data

Our k-mer based PBM model, PLAR-PBM (Figure 3.1), is motivated by the biophysics of TF binding to the probes in PBM experiments. Following Zhao et al. [204], we denote by Y_i the experimentally measured intensity of the i-th probe on the PBM array. We denote by $F(i)$ the (unobserved) binding probability of the TF to this probe. While these two quantities are related, due to experiential errors and scaling they are not identical. We thus assume a simple linear model for the mapping between the two:

$$Y_i = a + cF(i) + \epsilon_i \tag{3.1}$$

where ϵ_i is the error term and a and c are scaling factors. Since each probe is much longer than the motif itself (probe length is 36bp while motifs are often between 6-10bp) we follow BEEML-PBM [204], and express the binding probability $F(i)$ as the sum of the binding probabilities over all possible motifs encoded by the probe. Let k be the length of a TF binding site and L be the length of the variable region on the probe, we have:

$$F(i) = \sum_{j=1}^{L-k+1} \lambda_j \cdot \beta_{S_i(j)} \tag{3.2}$$

where λ_j is the position effect at position j (see below for the details on the position effect) and $\beta_{S_i(j)}$ is the binding probability to $S_i(j)$, the k-mer at the j-th position of the i-th probe. The term β_s is symmetric for any k-mer s, i.e. $\beta_s = \beta_{\bar{s}}$, where $\bar{s}$ is the reverse complement of s.

Plugging in the equation of $F(i)$ into the linear model, we have:

$$Y_i = a + c\left(\sum_{j=1}^{L-k+1} \beta_{S_i(j)} \cdot \lambda_j\right) + \epsilon_i. \tag{3.3}$$

We can rewrite this model as:

$$Y_i = a + c\left(\sum_{s\in\Sigma^k} \beta_s X_{is}\right) + \epsilon_i \tag{3.4}$$

where Σ^k denotes all k-mers, and X_{is} is the number of times the k-mer s occurs in the i-th probe (weighted by the position effects):

$$X_{is} = \sum_{j:S_i(j)=s} \lambda_j \tag{3.5}$$

($X_{is} = 0$ if s does not occur in probe i).

Note that c and the β's are coupled in the above model and so they can not be estimated individually. Thus we can write the equation as a linear model:

$$Y_i = \beta_0 + \sum_{s\in\Sigma^k} \beta_s X_{is} + \epsilon_i \tag{3.6}$$

subject to the constraints that (1) $\beta_s \geq 0$ for any s; and (2) β_s is symmetric, resulting in $4^k/2$ parameters that we need to learn.

Our model differs from the PWM models of BEEML-PBM [204] since PWMs cannot capture the possibility of secondary motifs and the possible dependency of nucleotides at different positions of a motif. In our model, no such constraint is imposed, and we deal with the model complexity problem (too many parameters) through sparse linear regression (see below). However, both models use similar assumptions regarding the biophysical nature of PBM measurements and how these relate to k-mer binding probabilities. Thus our model enjoys the same benefits as those of the BEEML-PBM model: the parameters for the k-mers have clear meanings, and certain experimental artifacts (e.g. handling biases due to position and background effects) can be naturally incorporated (see below).

3.2.2 Position and background effects in modeling PBM data

Zhao et al. [204] considered several experimental artifacts that should be removed in order to improve the accuracy of PBM data. These include the position effect and the background effect that we adopted in our model, described briefly below. We refer to [204] for more details.

Berger et al. [17] observed that the position along a probe at which the TF binds affects the binding strength and thereby the fluorescence intensity. Zhao et al. [204] used a position effect term to explicitly model this effect. The position effect of the j-th position along a probe, denoted as λ_j, is defined as [204]

$$\lambda_j = \frac{\left\langle \frac{I_{\text{avg}}(S_{i,j})}{I_{\text{avg}}(S_i)} \right\rangle_n}{\sum_{m=1}^{L} \left\langle \frac{I_{\text{avg}}(S_{i,m})}{I_{\text{avg}}(S_i)} \right\rangle_n} \tag{3.7}$$

in which $I_{\text{avg}}(S_{i,j})$ denotes the average intensities of all probes containing k-mer S_i at position j, $I_{\text{avg}}(S_i)$ denotes the average intensities of all probes containing S_i at any position, and $<>_n$ denotes the average over the top n k-mers with the highest median intensities ($n = 25$).

Zhao et al. [204] observed that in a typical PBM experiment, only a small fraction of the probes have high intensities due to TF binding, and most other probes have low intensities due to background hybridization. They estimated the distribution of background intensities from the lower half of the binned distribution of all fluorescent intensities, and then the i-th probe is weighted by

$$W_i = \frac{O_i - B_i}{O_i} \tag{3.8}$$

in which O_i and B_i denote the observed and expected number of probes in the corresponding bin. We similarly weigh each probe by W_i in the regression model.

3.2.3 Learning the parameters of the linear model

The above linear model has approximately $4^k/2$ parameters (one for each k-mer and its reverse complement). To avoid overfitting, we use the lasso regression [43] to estimate the coefficients.

Lasso is a widely used approach to linear regression that encourages a sparse model where most of the coefficients are zero. In our problem, we have the additional requirement that the coefficients must be non-negative, and this is known as positive lasso [43]. We implemented the positive lasso by modifying the lars package in R according to [43]. To estimate the tuning parameter for the regularization terms, we used 5-fold cross validation. In order to enforce sparsity, we chose the maximum tuning parameter such that the training error of the model is within one standard deviation of the minimum training error achievable.

After learning the model parameters using positive lasso, we set $\beta'_s = \beta_s / \max_s \beta_s$ to be the binding probability of the TF to the k-mer s up to a scaling constant.

Since we do not know the width of the motif bound by each TF, PLAR-PBM searches for k-mers of different lengths. In order to speed up the calculation, we first run positive lasso using all short 4-6 mers. To allow longer k-mers to be considered, after the first run, all pairs from the top 100 such k-mers (based on regression coefficients) are tested to see if the prefix of one matches the suffix of the other, yielding longer $(k + 1)$-mers. This process is repeated until up to 8-mers have been added to the feature set. In addition, we also allow for gapped k-mers to be considered (see below).

3.2.4 Allowing gapped and longer k-mers

Since we do not know the width of the motif bound by each TF, PLAR-PBM searches for k-mers of different lengths. In order to speed up the calculation, we first run positive lasso using all short 4-6 mers. To allow longer k-mers to be considered, after the first run, all pairs from the top 100 such k-mers (based on regression coefficients) are tested to see if the prefix of one matches the suffix of the other, yielding longer $(k + 1)$-mers. In addition, to allow gapped k-mers to be considered, we look at all pairs of 4- and 5-mers. If any pair of such k-mers, after connected by up to 3 gaps, appear more than once in the top $1,000$ probes with highest intensities, we add this gapped k-mer to the potential feature set as well. The entire process is repeated until up to

8-mers (not counting gaps) have been considered to be included in the feature set.

3.2.5 PBM data

The PBM data for 284 mouse TFs were downloaded from UniPROBE[126] at http://the_brain.bwh.harvard.edu/uniprobe/ (see Appendix C for a full list). For many TFs, the PBM data contains two versions of microarrays that differ in their probe designs. In such cases, for simplicity, we only run our method on the first version denoted by "v1" in the corresponding PBM dataset.

3.2.6 Predicting TF binding to any sequences

Our model is trained on sequences of 36bp in length (the length of the variable region of probes in PBM experiments), however, in practice, we often need to predict TF binding to longer sequences, e.g. promoter regions up to thousands of base pairs long. To predict the binding of a TF to a longer sequence (Figure 3.1b), we first define the binding probabilities of the TF to each overlapping 36bp region ("*site*", which is the same length as the variable region on the probe on which the coefficients were learned) in that sequence:

$$B = \frac{1}{B_{\text{max}}} \sum_{s \in \Sigma^k} \beta_s C_s \tag{3.9}$$

in which β_s is the binding probability to the k-mer s learned by the regression model, C_s is the number of times that s occurs in this site, and B_{max} is the highest possible unscaled binding probability of any 36-mer that can be achieved for the TF and is used as a scaling constant. The interpretation of this equation is that the binding probability to a 36bp sequence is the sum of binding probabilities to each of the k-mer of the 36bp sequence [50, 66]. In practice, B_{max} is estimated, for each TF, from the highest unscaled binding probabilities to 100,000 randomly sampled 36bp sites. Then, the binding probability to the entire sequence is defined as the highest binding probability to any 36bp site in that sequence.

3.3 Method: Integrated model of TF binding *in vivo*

PBM experiments measure TF binding *in vitro*. *In vivo* binding depends on several factors including the cellular environment and the chromatin state of the bound region. In addition, it has been shown that functional TFBSs tend to be evolutionary constrained [173, 195]. In this section, we describe integrated models that integrates our PBM motif learning and scanning method with these additional data sources in order to determine tissue specific binding.

3.3.1 Incorporating DNase I hypersensitivity data

Let B_i be the probability of binding of the TF of interest to a 36bp genomic region (a *site*) indexed by i based on the PBM model (Equation 3.9), reflecting the potential of TF binding *in vitro*. We are interested in the *in vivo* occupancy of the site, denoted as X_i. We assume that X_i is influenced by the chromatin state, which can be represented as a simple binary indicator variable, A_i (it is 1 if the chromatin is open/accessible and 0 otherwise). When the chromatin is open ($A_i = 1$), the occupancy X_i equals B_i; whereas a closed chromatin at that location means that $X_i = 0$. Thus, X_i is simply the product of B_i and $P(A_i = 1)$:

$$X_i = B_i \cdot P(A_i = 1) \tag{3.10}$$

The chromatin state variable can be partially determined using experimental data. Here we use DNase I hypersensitivity (HS) data (Figure 3.1c) which is available for several mouse tissues.

The DNase I hypersensitivity data for 55 mouse tissue/cell types are downloaded from the mouse ENCODE project website at `http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDnase/`[118] (see Appendix D for a full list). For each tissue/cell type, the “Signal” track that represents the normalized tag density is used. Chromatin in genomic regions with higher tag density are more open and therefore more accessible. For each input sequence (600bp genomic region), the tag density in the corresponding tissue at all positions are extracted, and for each overlapping 36bp window in the sequence, the max density is used as the density measurement for that window. For tissue/cell types where multiple replicates are available, the median values of all replicates are used.

The DNase I tag densities (integer data) are modeled using a mixture of negative binomial distributions similar to [139]: each region has probability π_1 to be open ($\pi_1 = P(A_i = 1)$), and probability $1 - \pi_1$ otherwise. The tag densities of regions follow two different negative binomial distributions depending on whether $A_i = 1$ or not. To estimate the model parameters, we sample 100,000 random 36bp regions from the upstream 10kb of all genes in the data. Parameters are estimated by maximizing the likelihood function:

$$L(\mathbf{r}, \mathbf{p}, \pi_1) = P(D|\mathbf{r}, \mathbf{p}, \pi_1) = \prod_{i=1}^{N} [(1 - \pi_1) \cdot NB(D_i|r_0, p_0) + \pi_1 \cdot NB(D_i|r_1, p_1)] \tag{3.11}$$

in which $D = (D_1, \cdots, D_N)$ denotes the DNase I tag densities of the sampled regions, and $r_k, p_k (k = 0, 1)$ are the parameters of the negative binomial distributions of the two components. After the parameters $\phi = \{\pi_1, r_0, p_0, r_1, p_1\}$ are estimated, the probability of each site being open can be calculated by

$$P(A_i = 1|D_i) = \frac{\pi_1 \cdot NB(D_i|r_1, p_1)}{(1 - \pi_1) \cdot NB(D_i|r_0, p_0) + \pi_1 \cdot NB(D_i|r_1, p_1)}. \tag{3.12}$$

3.3.2 The full integrated model

To further incorporate the conservation data into the integrative model, we consider the following graphical model:

$$X_i \leftarrow Z_i \rightarrow C_i \rightarrow S_i \tag{3.13}$$

in which X_i is the score of the site i that combines PBM and DNase data as described in the previous section, Z_i is a binary variable indicating whether the site i is a true binding site *in vivo* or not (hidden variable), C_i is a binary variable indicating whether the site i is conserved or not (hidden), and S_i is a measure of the evolutionary conservation of the site. The model assumes that true TFBSs have a higher occupancy score. Similarly, when $Z_i = 1$, C_i is more likely to be 1 as well (a true binding site is more likely to be conserved), and this is reflected by a higher conservation score S_i. The goal is to infer Z_i from the observed data X_i and S_i. The evolutionary conservation measure we used is the PhastCons scores [167] (phastCons 46way vertebrates) downloaded from the the UCSC Genome Browser (`http://genome.ucsc.edu`). For each 36bp segment, we chose the max score over the 36bp window as the PhastCons score of the sequence itself.

We model the first part ($X_i \leftarrow Z_i$) by a Beta distribution:

$$P(X_i|Z_i = k) \sim \text{Beta}(\nu_k \rho_k, \nu_k(1 - \rho_k)), k = 0, 1 \tag{3.14}$$

where ρ_k is the mean of $X_i|Z_i = k$ and ν_k is the pseudocount for Beta distribution.

For the second part ($Z_i \rightarrow C_i \rightarrow S_i$), we define $\alpha_1 = P(C_i = 1|Z_i = 1)$ as the fraction of conserved sites among true TFBSs, and similarly $\alpha_0 = P(C_i = 1|Z_i = 0)$ as the fraction of conserved sites among non-binding sites. The conditional distribution of S_i given Z_i is:

$$\begin{aligned}
P(S_i|Z_i = 1) &= P(S_i, C_i = 1|Z_i = 1) + P(S_i, C_i = 0|Z_i = 1) && (3.15)\\
&= P(C_i = 1|Z_i = 1)P(S_i|C_i = 1) + P(C_i = 0|Z_i = 1)P(S_i|C_i = 0) && (3.16)\\
&= \alpha_1 P(S_i|C_i = 1) + (1 - \alpha_1)P(S_i|C_i = 0) && (3.17)\\
&= \alpha_1 \frac{P(C_i = 1|S_i)P(S_i)}{P(C_i = 1)} + (1 - \alpha_1)\frac{P(C_i = 0|S_i)P(S_i)}{P(C_i = 0)} && (3.18)\\
&= P(S_i)\left[\alpha_1 \frac{S_i}{\gamma} + (1 - \alpha_1)\frac{1 - S_i}{1 - \gamma}\right] && (3.19)
\end{aligned}$$

where Equation 3.19 follows because according to the definition of phastCons score, $P(C_i = 1|S_i) = S_i$, and $\gamma = P(C_i = 1)$ is the total fraction of conserved sequences in the genome which can be estimated simply as the average of S_i:

$$\begin{aligned}
\gamma &= P(C_i = 1) && (3.20)\\
&= \sum_{S_i} P(C_i = 1|S_i)P(S_i) && (3.21)\\
&= \sum_{S_i} S_i P(S_i) && (3.22)\\
&= E(S_i) && (3.23)
\end{aligned}$$

Similarly,

$$P(S_i|Z_i = 0) = P(S_i)\left[\alpha_0 \frac{S_i}{\gamma} + (1 - \alpha_0)\frac{1 - S_i}{1 - \gamma}\right] \tag{3.24}$$

Let $p = P(Z_i = 1)$ be the fraction of the functional TFBSs, the likelihood function over the entire dataset is given by:

$$P(X, S|\theta) \propto \prod_i [p \cdot \text{Beta}(X_i|\rho_1, \nu_1)P(S_i|Z_i = 1) + (1 - p) \cdot \text{Beta}(X_i|\rho_0, \nu_0)P(S_i|Z_i = 0)] \tag{3.25}$$

where $\theta = (\rho_1, \nu_1, \rho_0, \nu_0, \alpha_1, \alpha_0, p)$ represents the model parameters. Note that the terms $P(S_i)$ are not shown because they are independent of model parameters. We estimate θ by maximizing the likelihood function for each TF and tissue/cell type on 100,000 randomly sampled sites from gene promoters. The vast majority of sequences are not bound by a TF, so we could use all the data to fit ρ_0, ν_0 and α_0 and then fix them. Thus only $(\rho_1, \nu_1, \alpha_1, p)$ are the free parameters to be estimated.

After the parameters are estimated, given any new site, its probability of being bound is predicted by

$$\frac{P(Z_i = 1|S_i, X_i, \hat{\theta})}{P(Z_i = 0|S_i, X_i, \hat{\theta})} = \frac{P(X_i|Z_i = 1, \hat{\theta})}{P(X_i|Z_i = 0, \hat{\theta})} \cdot \frac{P(S_i|Z_i = 1, \hat{\theta})}{P(S_i|Z_i = 0, \hat{\theta})}. \tag{3.26}$$

3.4 Method: Comparison with other methods

To evaluate the performance of our method, we obtained ChIP-seq data for 8 TFs for which PBM data and tissue specific DNase I hypersensitivity data were available (Table 3.1). For each dataset, the top 3000 peaks with highest enrichment are extracted, and the 600bp genomic regions centered on the reported peaks are used as the positive sequences bound by the TF. Then, 600bp sequences that (1) are upstream of and (2) 300bp apart from each positive sequence, and (3) do not overlap with any other positive sequences, are used as negative sequences. For each sequence to be tested, the binding probabilities of the TF to each overlapping 36bp window (*site*) in that sequence is calculated based on the model learned (see Section 3.2.6), and the maximum probability is defined as the binding probability to that sequence.

Table 3.1: List of TFs and ChIP-seq experiments used in evaluation

TF in ChIP-seq	TF in PBM	Tissue/Cell type	GEO ID	Reference
Esrrb	Esrra	Embryonic stem (ES) cells	GSM288355	[27]
Klf4	Klf7	ES cell	GSM288354	[27]
Sox2	Sox12	ES cell	GSM288347	[27]
Oct4	Pou2f1	ES cell	GSM288346	[27]
FoxA2	FoxA2	Liver	GSE25836	[174]
Crx	Crx	Retina	GSM499736	[32]
Nkx2-5	Nkx2-5	Heart	GSM558906	[65]
Srf	Srf	Heart	GSM558907	[65]

We performed a comprehensive comparison of PLAR-PBM with several other methods that could be or have been used in predicting TF binding on real sequences, using area under the ROC curve (AUC) as the criteria. A detailed description of all methods is provided below.

3.4.1 Methods that only use PBM data

1. PWM-based methods (including PBM PWM and BEEML PWM). The PWMs are converted to a log-odd scoring matrix using a zeroth-order background estimated from each input sequence respectively, and then all overlapping k-mers (k being the width of the PWM) on both strands are scored by the log-odd scoring matrix. The maximum score is used as the score for the sequence. In cases where a secondary PWM was reported, only the primary one was considered.

2. BEEML Energy. The binding energy matrices reported by BEEML-PBM [204] are used to score each overlapping k-mers (k being the width of the energy matrix). The score for a k-mer is the summation of the binding energies of the corresponding bases at each

position. The minimum energy is used as the score for the sequence.

3. Max E-score. The E-score [17] for each k-mer in the sequence is considered and the maximum E-score is used as the score for the sequence.

4. Occupancy score. The background-subtracted median intensities of all k-mers with E-scores higher than 0.35 is summed and the sum is used as the score for the sequence [207].

5. SVR. The SVR-based method [1] is run on overlapping 36bp windows in each input sequences and the maximum score is reported as the score for that sequence. The number of k-mers used is set to 1000 to speed the calculations, and default values are used for other parameters ($k = 13$ and $m = 5$).

3.4.2 Methods that combine sequence with DNase data

Neph et al. [124] predicted tissue-specific TF binding sites in human by overlapping motif scanning results with DNase I footprints, the actual locations bound by TFs within DNase I hypersensitive sites that are protected from DNase I cleavage [55]. In order to compare our method with this under a similar framework, we applied a similar approach by first using FIMO ([57], in the MEME suite v4.8.1) to scan the sequences for occurrences of the PWMs reported with the PBM data, and then overlapping the motif occurrences (p-value $< 1e-4$) with DNase I hypersensitivity data. When a secondary PWM is available, both PWMs are used to scan the sequences. Default parameters are used for FIMO. AUC is obtained by varying the cutoff on the DNase data.

3.5 Method: Identifying tissue-specific TF activities

We use the PBM predictions (sequence-specific, but common in all tissues) as well as the tissue-specific DNase I HS data to identify TFs likely to be active in each tissue. Intuitively, if a TF f is active in a tissue T, then the binding sites of f should be overrepresented in the open chromatin regions of T. To quantify this overrepresentation, we define $R(f, T)$ as the fraction of DNase hypersensitive sites in tissue T that contain high-scoring binding sites of f. The high scoring sites are defined as those that have binding probabilities (according to the PBM model, as defined in Equation 3.9) higher than the top 0.1% of the binding probabilities for all possible sites for that TF. In practice the binding probability distribution of a TF is estimated from the 100,000 sampled sites. For each tissue, the open sites are defined as sites within the promoter regions of all genes with DNase tag densities higher than 15 (so that $P(A_i = 1|D_i)$ is close to 1). Sites that are constitutively open in more than 1/3 of all tissues considered are excluded from the counting.

The activity score of a TF f in tissue T is defined as:

$$Activity(f, T) = \frac{R(f, T)}{R(f, \bar{T})} \tag{3.27}$$

where $\bar{T}$ denotes all tissues other than T. This is used as a measure of the likely activity of the TF in that tissue.

3.6 Results

An overview of our method is shown in Figure 3.1. Starting with raw fluorescent intensities measured by PBM that represents binding strength of the TF of interest to each probe on the array, we first infer binding probabilities to each individual k-mers with a biophysically-motivated model PLAR-PBM (Figure 3.1a), and such information based on PBM alone can be used to score a new sequence for potential TFBS (Figure 3.1b). To predict tissue-specific TF binding, we model DNase I hypersensitivity data for each tissue that represents chromatin accessibility (Figure 3.1c, d), and combine such information, possibly together with sequence conservation and other types of data, with the inferred binding probabilities mentioned above to predict *in vivo* binding sites (Figure 3.1e). See below and also Methods for details.

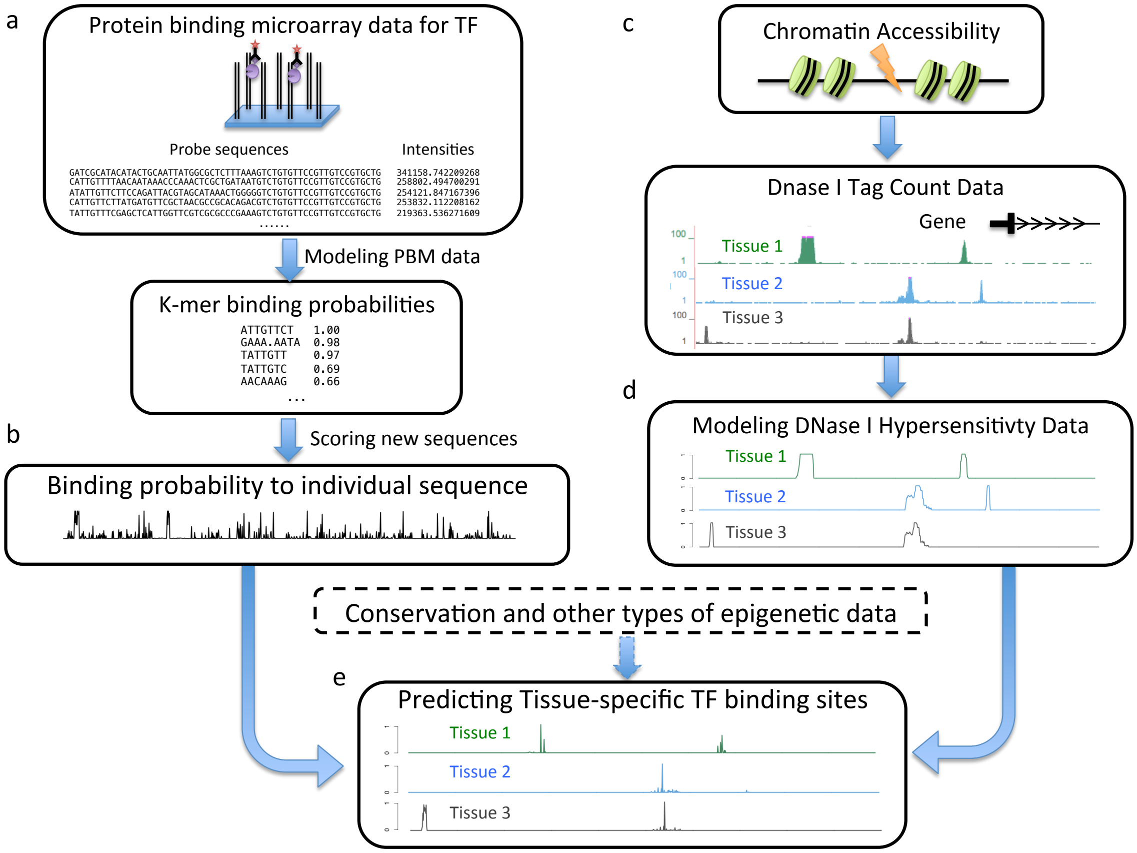
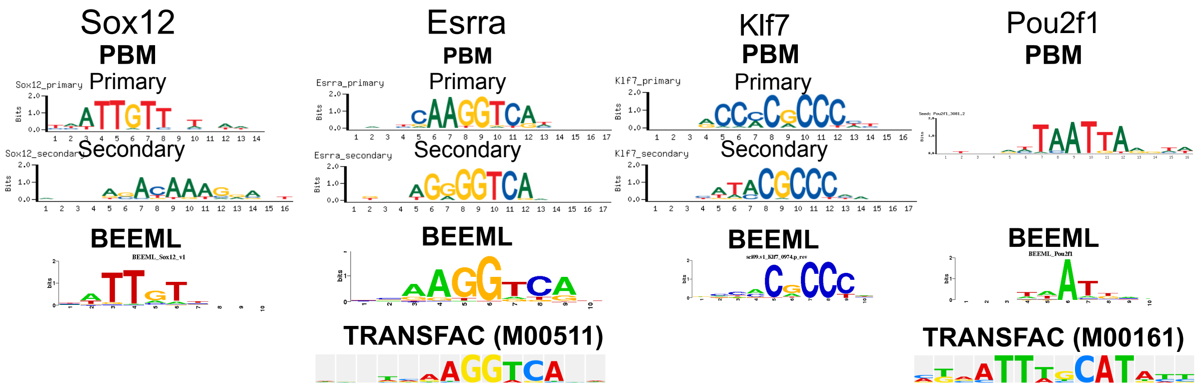

a
Protein binding microarray data for TF
Probe sequences
Intensities
GATCGCATACATACTGCAATTATGGCGCTCTTTAAAGTCTGTGTTCCGTTGTCCGTGCTG 341158.742209268
CATTGTTTTAACAATAAACCCAAACTCGCTGATAATGTCTGTGTTCCGTTGTCCGTGCTG 258802.494700291
ATATTGTTCTTCCAGATTACGTAGCATAAACTGGGGGTCTGTGTTCCGTTGTCCGTGCTG 254121.847167396
CATTGTTCTTATGATGTTCGCTAACGCCGCACAGACGTCTGTGTTCCGTTGTCCGTGCTG 253832.112208162
TATTGTTTCGAGCTCATTGGTTCGTCGCGCCCGAAAGTCTGTGTTCCGTTGTCCGTGCTG 219363.536271609
......
Modeling PBM data
K-mer binding probabilities
ATTGTTCT 1.00
GAAA.AATA 0.98
TATTGTT 0.97
TATTGTC 0.69
AACAAAG 0.66
...
b
Scoring new sequences
Binding probability to individual sequence
c
Chromatin Accessibility
Dnase I Tag Count Data
Gene
100
Tissue 1
1
100
Tissue 2
1
100
Tissue 3
1
d
Modeling DNase I Hypersensitivty Data
Tissue 1
Tissue 2
Tissue 3
Conservation and other types of epigenetic data
e
Predicting Tissue-specific TF binding sites
Tissue 1
Tissue 2
Tissue 3

Figure 3.1: Overview of our method. (a,b) Starting with raw PBM data for a TF represented as fluorescence intensities to each individual probes, we first infer binding probabilities to individual short k-mers, and then a given sequence can be scored by such inferred binding probabilities. (c,d) To predict *in vivo* TF binding, we take as input the tissue-specific DNase I hypersensitivity data (tag counts) and convert them to probabilities that represent chromatin accessibility for each position in the genome at each tissue/cell types. (e) The PBM data, DNase data and other types of data including sequence conservation are combined using an integrative model. See Section 3.4 for details.

3.6.1 The biophysically-motivated PBM model accurately infers binding specificities of TFs

We developed a biophysically-motivated k-mer based model, named PLAR-PBM, to infer the binding profiles of a TF from protein binding microarray (PBM) data (Figure 3.1a). The model combines the benefits of recent PWM-based biophysical methods (for example, BEEML-PBM [204]) with the ability of PBMs to capture dependencies between positions in a given motif. The parameters of our model have well-defined meanings, and it naturally accounts for the experimental artifacts such as positional biases (Section 3.2.2). On the other hand, it avoids the independence assumption made by BEEML-PBM that each position of the binding site contributes additively to the total binding energy of the site. This allows a richer representation of the binding specificities of TFs. In addition, since we do not rely on PWMs, there is no need to pre-specify the motif length, which makes the model much more general. We use lasso regression to learn model parameters that represent binding probabilities to individual k-mers, where k is determined in the learning procedure. This results in a sparse model with relatively few k-mers having nonzero binding probabilities.

We illustrate the results of PLAR-PBM using four TFs including Sox12, Esrra, Klf7 and Pou2f1. Figure 3.2 presents the PWMs derived from the PBM data for these TFs by the Seed-and-Wobble algorithm [17] (denoted as PBM PWMs) and the BEEML-PBM method [204] (denoted as BEEML PWMs), PWMs in TRANSFAC [111] for the corresponding TFs when available, as well as all the k-mers estimated by PLAR-PBM that have binding probabilities above 0.5. For PBM PWMs, when a secondary binding preference was derived [6], both the primary and secondary PWMs are shown. For Sox12, the learned k-mers match well with both the primary and secondary PBM PWMs (Figure 3.2a). For Esrra and Klf7, k-mers matching the consensus sequences of their primary PBM PWMs respectively are all predicted to have high binding probabilities (Figure 3.2b and c). But there are subtle yet important differences between the PWMs and the predicted k-mers. Some of the top k-mers of Esrra and Klf7 are different from the consensus



ATTGTTCT	1.00	CAAGGTTA	1.00	CCCCGCCC	1.00	ATTTGCAT	1.00
GAAA.AATA	0.98	CAAGGGCA	0.93	CCACGCCC	1.00	ATTAGCAT	0.92
TATTGTT	0.97	CAATGTCA	0.82	CCCCACCC	0.68	ATTTACAT	0.79
TATTGTC	0.69	CAAGGTCG	0.77	GCCACACC	0.64	CTCATTA	0.68
AACAAAG	0.66	CAAGGTCA	0.76	CCACACCC	0.64	TTAGCATA	0.59
AAACAATG	0.62	AAGGTCA	0.67	CCACGCCT	0.51	TTTACATA	0.55
ATTGTTC	0.62	CAAGGTGA	0.65			ATTTAAAT	0.52
ATGTTA	0.57	CAAGGACA	0.54			ATTAACAT	0.51
ATTGTCTT	0.50						

Figure 3.2: Top inferred k-mer binding probabilities for (a) Sox12, (b) Esrra, (c) Klf7 and (d) Pou2f1. The PBM PWMs, BEEML PWMs and TRANSFAC motifs (when available) for these four factors are also shown. k-mers are colored according to whether they match the consensus sequences of the primary PBM PWM (red), secondary PBM PWM (blue) or the TRANSFAC motif (green).

sites in the PWMs in one or two highly-specific positions, yet have strong predicted binding. For example, the second strongest k-mer of Esrra, CAAGGGCA, is predicted to have a binding probability of 0.93, and it also had a high associated E-score [17]; yet it does not match either PBM PWM at the 6th position. Such k-mers, identified by PLAR-PBM but missed by the PWMs, may have important influences on TFBS predictions. Furthermore, PLAR-PBM provides important quantitative information about binding. For example, for both Esrra and Klf7, k-mers matching the consensus sequences of the secondary PBM PWMs respectively only have predicted binding probabilities ranging from 0.2 to 0.44, much less than those matching the primary PWM. In the case of Pou2f1, none of the inferred top k-mers match the PBM PWM, and the BEEML PWM is not very specific (Figure 3.2d). However, many of these top k-mers closely match the consensus sequence of the TRANSFAC motif for Pou2f1 derived from literature evidence (Figure 3.2d). Overall, the binding probabilities to individual short oligonucleotides that PLAR-PBM inferred from PBM data are largely consistent with known motifs for the corresponding factors, and they provide a more reliable and high-resolution representation of TF binding preferences compared with PWM representations.

3.6.2 PLAR-PBM outperforms other methods in predicting *in vivo* and *in vitro* TF binding from PBM data

We next used the inferred binding probabilities to predict *in vivo* TF binding. We collected 8 published mouse ChIP-seq datasets for which the PBM data for the same TF or for a TF with a similar DNA-binding domain is available (Table 3.1). From each ChIP-seq dataset, the top 3000 peaks with highest enrichment are extracted, and the 600bp genomic regions centered on the reported peaks are used as the positive sequences bound by the TF. Then, 600bp sequences that (1) are upstream of and (2) 300bp apart from each positive sequence, and (3) do not overlap with any other positive sequences, are used as negative sequences. We evaluated multiple methods by their abilities to correctly classify the two sets of sequences (see Methods).

We compared PLAR-PBM with several other methods that predict affinities of TF binding to given sequences. Three of these methods use PWMs or energy matrices derived from the PBM data to scan a given sequence. These include (1) the PBM PWM method which uses PWM derived from the PBM data by the Seed-and-Wobble algorithm [6, 17], (2) the BEEML Energy method which uses an energy matrix derived from the PBM data based on BEEML-PBM [204], a biophysical TF-DNA binding model, and (3) the BEEML PWM method which uses the PWM converted from the BEEML energy matrix. We also compared against methods that directly use the PBM data itself. These include the max E-score method [17] and the occupancy score method [207] which are based on a rank-based statistic called E-score derived from the PBM data. In addition, we also compared with a support vector regression (SVR) based method [1] which uses a novel string kernel to map sequences to intensities measured by PBMs (see Section 3.4 for details).

Performances of the methods are evaluated by area under the ROC curve (AUC) and are summarized in Figure 3.3. As can be seen, for 4 of the 8 TFs tested (Esrrb, Sox2, Oct4 and Crx) PLAR-PBM outperforms all other methods (Esrrb, p=1.10e-6; Sox2, p=1.98e-9; Pou2f1, p=1.49e-2; Crx, p=3.24e-3 comparing the AUC of PLAR-PBM against the second best method using one-sided DeLong test [39] implemented in the pROC package in R [150]). More generally, PLAR-PBM is always either the top or second in terms of performance. Importantly, in all cases where PLAR-PBM ranks the second, its AUC is still comparable to the highest AUC achieved by the top method for that TF and is usually significantly higher than the AUC by the 3rd method (see details in Table 3.2). As for the other methods, two of the methods that use PBM data directly (E-score and Occupancy score) perform well overall though in some cases have significantly lower AUC than PLAR-PBM (for example, Srf for occupancy score and Crx for E-score). PWM-based methods, in general, do not perform as well indicating that the dependency between nucleotides in the motif, which is ignored by PWMs, may be important for accurate identification of TF targets.

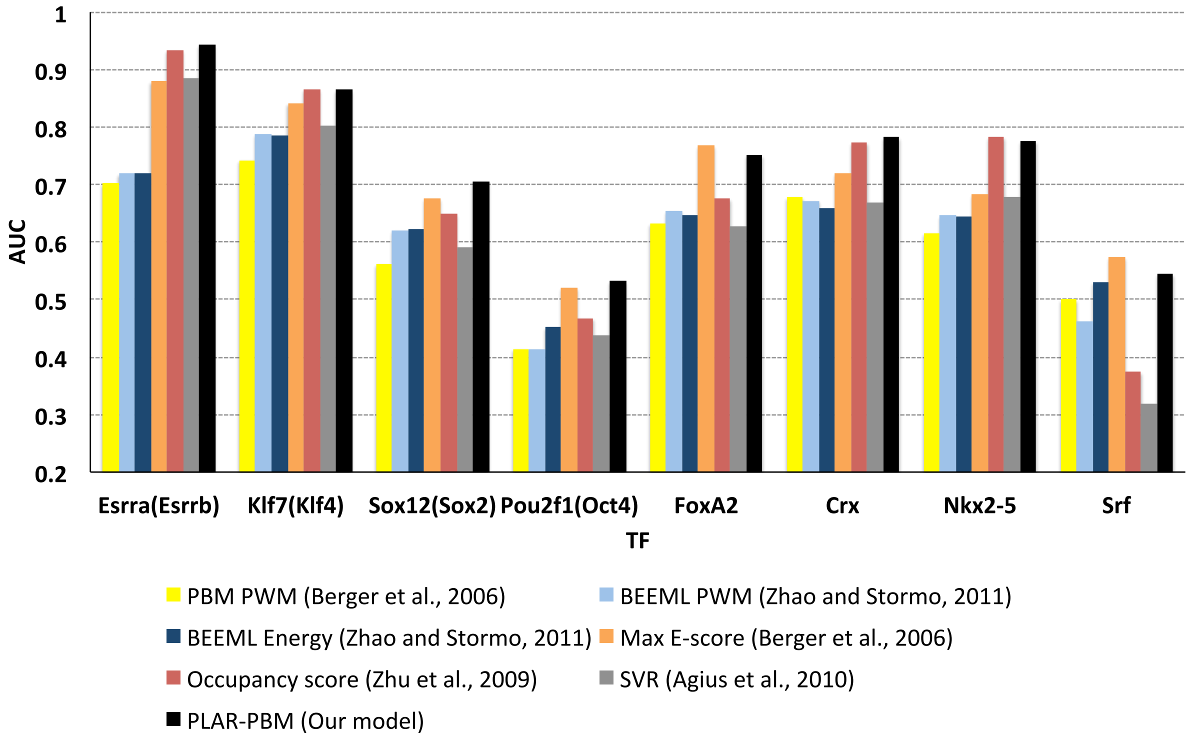


Figure 3.3: Area under ROC curve (AUC) of difference methods that use PBM data alone to predict *in vivo* binding sites. Shown on the x-axis are 8 TFs with ChIP-seq data available for the same TF or for a TF with a similar DNA-binding domain. In the latter case the PBM TF is shown in front and the ChIP-seq TF is shown in the parentheses.

Table 3.2: p-values of AUC comparisons for the 4 TFs for which PLAR-PBM ranks the 2nd

TF	1st method	AUC(1st)	p (1st>2nd)	AUC(PLAR)	3nd method	AUC(3rd)	p (2nd>3rd)
FoxA2	Max E	0.7716	8.58e-5	0.7519	Occ	0.6757	7.44e-53
Klf4	Occ	0.8658	0.360	0.8647	Max E	0.8519	1.09e-4
Nkx2-5	Occ	0.7829	4.31e-2	0.7760	Max E	0.6842	1.56e-82
Srf	Max E	0.5740	5.00e-4	0.5450	Occ	0.3744	9.75e-75

Max E: Max E-score;

Occ: Occupancy score;

AUC(1st): AUC of the 1st ranking method;

AUC(3rd): AUC of the 3rd ranking method;

p (1st vs 2nd): p-value testing whether the AUC of the 1st ranking method is higher than that of the 2nd ranking method (PLAR-PBM in all cases);

p (2nd vs 3rd): p-value testing whether the AUC of the 2nd ranking method (PLAR-PBM in all cases) is higher than that of the 3rd ranking method;

All p-values are calculated using one-sided DeLong test [39] implemented in the pROC package in R [150]

To further evaluate the ability of PLAR-PBM to predict *in vitro* TF binding, we applied PLAR-PBM on 98 mouse TFs with PBM data measured on a second array with a different probe design available, and investigated how accurately PLAR-PBM can predict the probe intensities on the alternative array. To evaluate its performance, we compared PLAR-PBM with the two best performing methods in the evaluation on *in vivo* data mentioned above, namely max E-score and occupancy score. The following two criteria is used: (1) the overall Pearson correlation coefficients of the predicted values and the original values for each probe; and (2) out of the top 100 probes on the alternative array with highest known intensities, how many are still predicted to have an intensity within the top 100 [1]. The first measure above provides an overview of how each method works in general, and the second measure above specifically focuses on probes with highest intensities which are more likely to be bound *bona fide* by the TF being studied. Figure 3.4 provides the result for these evaluations. As can be seen, PLAR-PBM clearly outperforms the max E-score method on 83% and 86% of the TFs studied under the two criteria respectively. For occupancy score, although PLAR-PBM ties with it in terms of the correlations, for 69% of the TFs PLAR-PBM correctly predicted more of the top 100 probes than occupancy score. Thus, PLAR-PBM also works well on reproducing *in vitro* binding data from PBM experiments.

3.6.3 Integrated model of PBM and DNase I hypersensitivity data significantly improves TFBS prediction accuracy

PBM data, although powerful, only measures *in vitro* binding. Therefore, even with sophisticated methods, the ability of using PBM data alone to predict *in vivo* binding is limited. DNase I hypersensitive (HS) sites are regions of chromatin that are very sensitive to DNaseI cleavage [59], and previous studies have shown that such hypersensitivity correlates with TF binding [73, 95]. To better predict tissue-specific *in vivo* binding sites, we developed a model for integrating DNase I HS data with PBM data. For each 36bp genomic region ("*site*"), we assume the chromatin of the site could exist in two states: open or closed, and only in the open state, the chromatin is

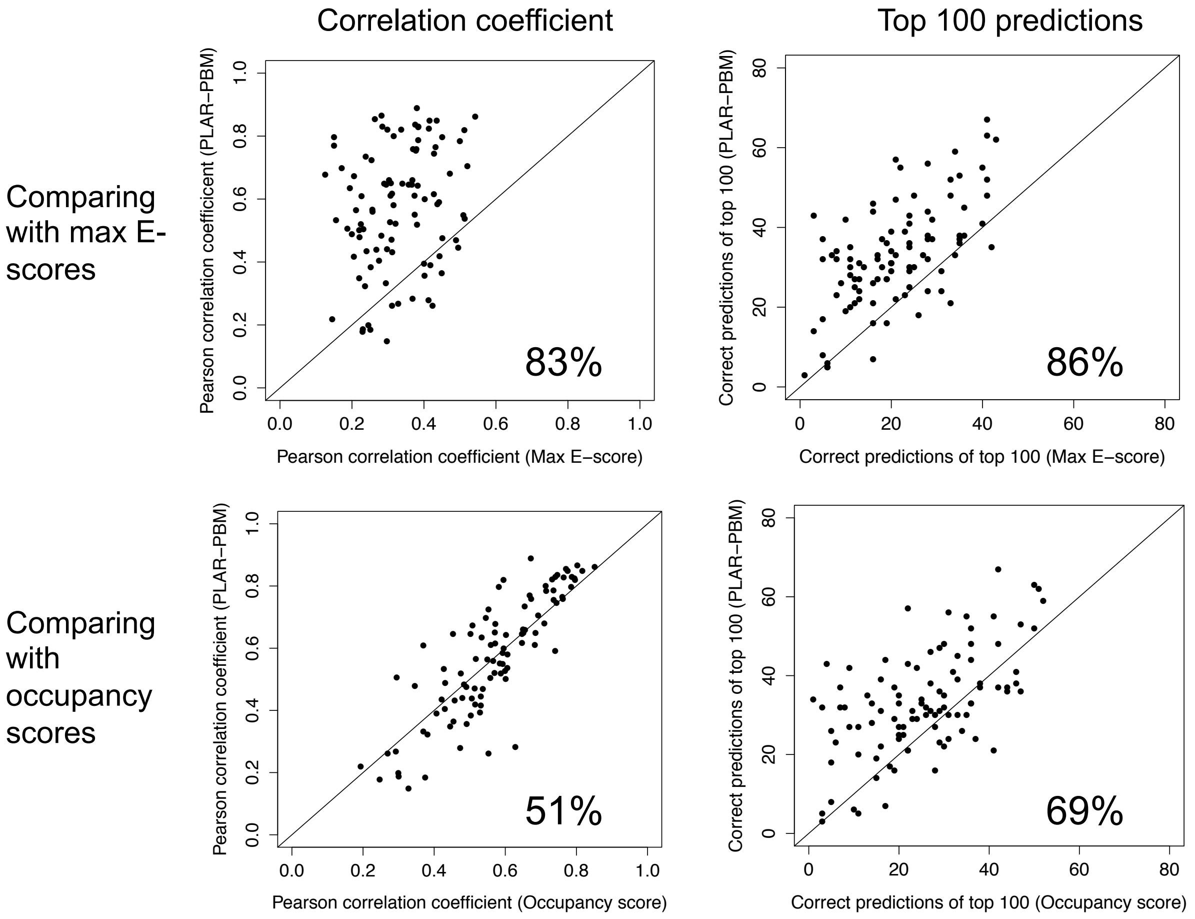


Figure 3.4: Comparing PLAR-PBM with max E-score (top) and occupancy scores (bottom) in predicting the intensities of PBM data for 98 mouse TFs with an alternative array design available. All methods are evaluated using two criteria: overall Pearson correlation coefficient (left) and out of the top 100 probes with highest known intensities, how many are still predicted to be within the top 100 (right) (see text for details). Each point represents the result of one TF. Percentages reflect the percent of TFs for which PLAR-PBM works better than the method being compared using the corresponding criteria.

accessible to binding by a TF molecule. We infer the chromatin state by using a mixture model for the DNase HS data: the open state should be associated with higher tag densities from the DNase data, and the closed state with lower densities. The *in vivo* occupancy of a site is then estimated as the probability of binding *in vitro* estimated from the PBM data as described in the previous section, multiplied by the probability that the site is in an open state inferred from the DNase HS data (see Section 3.3.1 for details).

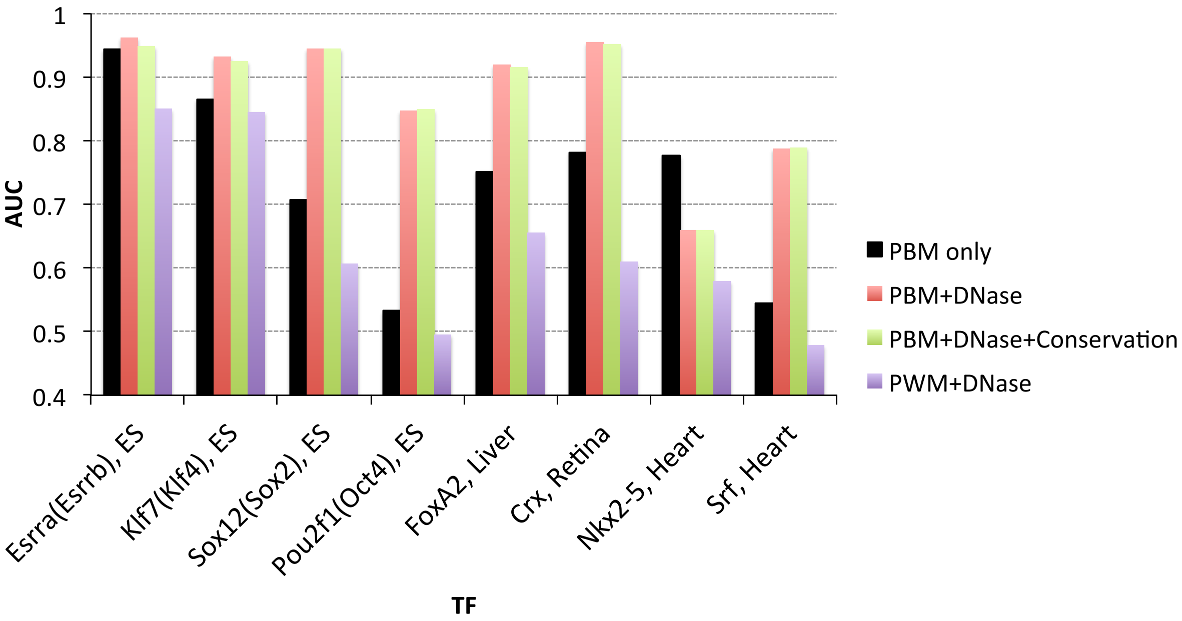


Figure 3.5: AUC of different methods that integrate PBM data or PWM scanning with DNase I hypersensitivity and/or sequence conservation data to predict *in vivo* TF binding.

Figure 3.5 presents the AUCs from applying the integrated model to predict *in vivo* TF binding in the corresponding tissues for the same 8 TFs studied in the previous section. Compared with using PBM data alone (black bars), the incorporation of DNase I HS data in the corresponding tissues (red bars) significantly improves performance for 7 of the 8 TFs. The biggest improvements are seen for TFs for which the results when using only PBM data are relatively poor. For example, when predicting Srf binding sites in heart, even the best methods analyzed above achieve an AUC only slightly greater than random (Figure 3.3). Combined the DNase data, the performance of our method is improved by 43% from 0.5450 to 0.7793 (Figure 3.5). Similar improvement is also observed for Oct4 (from 0.5325 to 0.8468).

While to the best of our knowledge no method has been used to predict global TF binding across a large number of tissues, a simple strategy for this purpose is to intersect sites that have high-scoring PWM matches for the TF with DNase HS sites. Such strategy has been used by several papers including Neph et al. [124]. We thus compare our approach with this simple intersection method by identifying sequence locations that lie in the overlap of high scoring PWM sites and high DNase I HS regions for the above 8 TFs (Section 3.4.2). As shown in Figure 3.5 (purple bars), this strategy, although intuitive and easy to use, lead to AUCs significantly lower than ours for all eight TFs. These results thus demonstrate that our integrated model using both PBM and DNase HS data is an effective approach to predicting TF binding sites *in vivo*.

In addition to DNase I hypersensitivity data, *bona fide* TF binding sites are usually under evolutionary pressure and therefore more conserved [173, 195]. In order to see whether conservation data could help in predicting *in vivo* binding sites, we further extended the above model to incorporate phastCons scores [167] for each site (see Section 3.3.2 for details). Performance of the full model that incorporates this additional information is shown in Figure 3.5 (green bars). As can be seen, while in some cases adding the conservation information very slightly improves performance (for example for Srf and Oct4), overall using conservation data does not lead to a significant improvement in prediction accuracy. Note however that our method is general and for other species with PBM data the use of sequence conservation may be more beneficial.

3.6.4 Combining PBM and DNase data enables the prediction of tissue-specific TF activities

The recently released mouse ENCODE project data provides DNase I hypersensitivity data for more than 50 mouse tissue/cell types (Section 3.3.1 and Appendix D). We set out to combine the PBM data for 284 mouse TFs in UniPROBE with such DNase data to predict tissue-specific TF targets and determine tissue-specific TF activities (Section 3.3).

Identifying TFs that are highly active in specific tissues is useful for determining the function of such TFs, and serves as an initial step for reconstructing the tissue-specific transcriptional regulatory networks. We predict how likely a TF is functional in any given tissue/cell type with an activity score for each TF-tissue pair. Our hypothesis is that if the TF is active in a tissue, it will bind a number of target sequences, thus the putative TF binding sites will be overrepresented in the DNase HS regions (see Section 3.3.1 for details). A higher activity score indicates that the TF is more active in the corresponding tissue (the expected value is 1 for non-active TFs). The complete results are provided on the Supplementary Website. In Figure 3.6a we illustrate these results by focusing on the activity scores calculated for 4 TFs (Gata3, Pou6f1, Crx and Hnf4a) across 18 representative tissue/cell types. Gata3 is known to function in mouse fetal liver haematopoiesis [133], and its expression had also been observed in leukemia cells [115]. Our results are in good agreement with the prior knowledge regarding Gata3's activity: the top two tissues predicted for Gata3 are E14.5 liver cells and the adult leukemia cell line. Similarly the top tissue for Pou6f1 is E14.5 whole brain, in agreement with its known role in brain development [201]. Crx is an important TF for regulating photoreceptor genes in retina [32, 67], and our method correctly determined that its activity score in that tissue is the highest. Finally, Hnf4a is a well known master regulator of liver- and kidney-specific genes [64, 102], as correctly predicted by our method. While we only show 18 tissues, for all four TFs the correct tissues shown in Figure 3.6a have the highest scores among all 55 tissues we tested (Supplementary Website).

To more globally validate these tissue-specific TF activities, we compared the correlation between our predicted TF activity scores and mRNA levels for the same TFs in the corresponding tissue (measured by qRT-PCR [142]) . Eight tissues and 222 TFs that are common to both datasets are used. Even though the two types of data (PBM and DNase vs. expression) measure completely different aspects of cellular activity, we observe a Pearson correlation coefficient of 0.229, which is highly statistically significant ($p < 10^{-8}$, permutation test, Figure 3.6b). Since

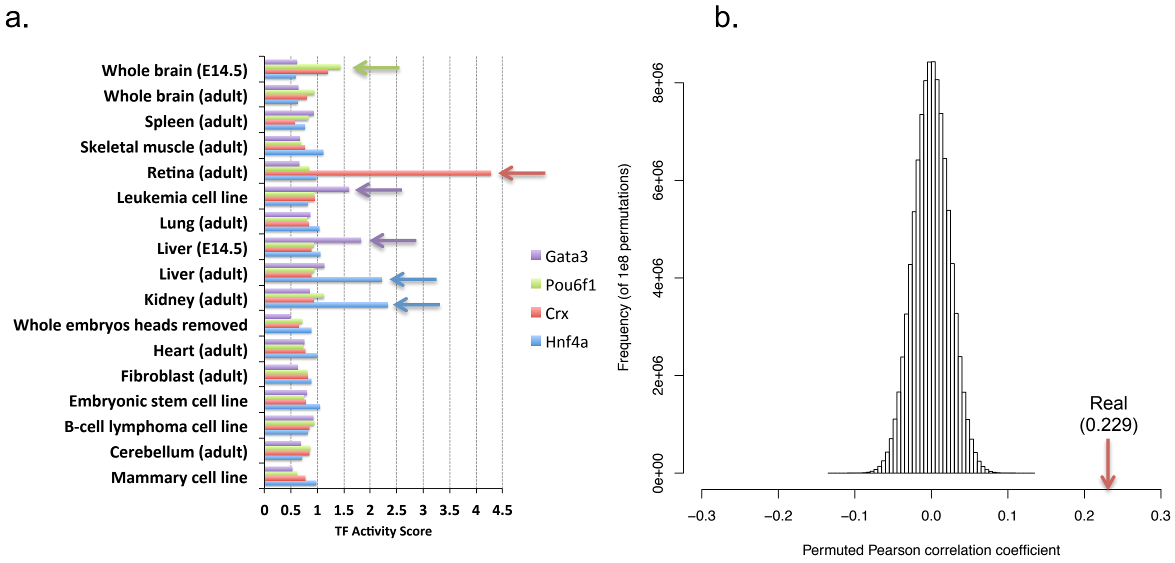


Figure 3.6: Results for tissue-specificity TF activity prediction. (a) Predicted tissue activity scores for 4 TFs across 18 representative tissue/cell types. Arrows indicate known functions of the TF in the corresponding tissue as supported by literature evidence. In all cases, the highest activity score matches a known tissue for the factor. (b) Pearson correlation coefficients between tissue specific expression experiments and the activity level predicted by our method. The distribution is based on 10^8 permutations of the activity scores. The value from the real predictions is indicated by the arrow on the right.

many TFs are only post-transcriptionally regulated, such a significant correlation provides strong support to the predictions computed by our method.

3.6.5 Existing literature strongly supports predicted TF activities in several tissues

To further validate our predictions and investigate their potentials to lead to new biological insights, we took a closer look at the TFs predicted to be active in the adult liver tissue. The top five such predictions are shown in Table 3.3A. Besides Hnf4a discussed above, Rara, Nr2f2, Rxra and Tcf7 are all known to either regulate liver-specific genes or are involved in maintaining liver metabolism and homeostasis (Table 3.3A). The $7th$ ranked factor Tcf7l2 (activity score of 1.38) was linked to type 2 diabetes risk in previous studies using SNP data [58], but the mechanism for its involvement was unclear. Our result indicates that it may have a regulatory role in liver metabolism. Indeed, a very recent study confirms its role in regulating key liver-specific metabolic genes[20]. Our result also assign a high liver activity score to Cutl1 (1.36, rank 8/284). Cutl1 was a known transcriptional repressor of terminal differentiation genes in several cell lineages including hepatocyte [153]. Recently, Cutl1 was identified as target of the liver-specific microRNA miR122 and a central mediator of the effects caused by the deregulation of miR122 in hepatocellular carcinoma [83]. Further down the list, Foxa2 (1.32, rank 10/284) is known to regulate lipid metabolism and ketogenesis related genes in liver [192], and Lef1 (1.31, rank 11/284) is a prognostic biomarker for liver metastasis in colorectal cancer. Other TFs ranked within the top 20 for liver include Tcf1 and Tcf2, members of the T-cell factor (Tcf) family that are critical for hepatocyte metabolism and function [128, 165]; Bhlhb2, which is involved in the regulation of lipogenesis in liver [71]; and Hmbox1, whose expression levels was shown to be reduced in liver cancer compared with surrounding normal tissues [37]. Overall, our predicted set of liver regulators is comprehensive, spanning several different classes of liver related activities including glucose and lipid metabolism and cancer, and including both repressors and activators.

In addition, Table 3.3B and C presents the top 5 predicted TFs for two more tissues (retina and B cell). As can be seen, for almost all of these TFs there is strong support for their tissue-specific activity in the predicted tissue.

3.6.6 Predicting genome-wide tissue-specific TF binding sites

We provide a resource for genome-wide tissue-specific TF binding sites predicted for 284 mouse TFs with PBM data available (Section 3.2.5) and 55 tissue/cell types with DNase data available (Section 3.3.1). A list of all the TFs and tissue/cell types is provided in Appendix C and D. To predict targets of TFs, the promoter regions (+/- 10kb around transcription start sites) for all mouse genes were scanned. The complete prediction results are available for download from the Supplementary Website at http://www.sb.cs.cmu.edu/PLAR-PBM.

3.7 Discussion

We presented a computational strategy which relies on a biophysically-motivated model, PLAR-PBM, to identify top scoring k-mers in PBM data. Our extensive evaluations demonstrate that using the selected k-mers to predict *in vivo* TF targets improves upon PWM-based methods suggested for this task, including methods that derived PWMs from the same PBM data used in our paper. This indicates that at least in some cases dependencies exist between bases in a motif and that using models that utilize such dependencies may improve the accuracy of the predicted targets. We next developed a strategy to integrate PBM, DNase HS data and sequence conservation data to provide the first comprehensive map of more than 200 TFs across 50 tissues. The combined model was shown to be highly accurate at predicting *in vivo* TF binding. Several of our predictions agree well with existing knowledge in literature regarding the role that some TFs play in specific tissues. The overall results are significantly correlated with independent gene expression data measured for these TFs across tissues even though such expression data was not

Table 3.3: Top five predicted TFs for liver, retina and B cell

TF	Score	Known functions in the corresponding tissue
A. Liver		
Hnf4a	2.22	Essential for maintaining hepatic gene expression and lipid homeostasis [64]
Rara	1.90	Important in maintaining liver homeostasis, and its disruption is linked to hepatocarcinogenesis [81]
Nr2f2	1.56	Expressed in liver, and known to regulate liver-specific genes [202]
Rxra	1.45	Important role in liver metabolism [186]
Tcf7	1.44	Downstream regulator in Wnt signaling which is critical in liver physiology and pathology [182]
B. Retina		
Crx	4.28	Regulates photoreceptor gene expression [32]
Pitx3	4.26	Required for normal retina formation in Xenopus and zebrafish [82, 164]
E2F3	4.06	Involved in retina progenitor cell development [26]
Pitx2	3.92	Pitx2-deficient mouse exhibits ocular abnormalities [54]
Gsc	3.89	Unknown function in retina.
C. CD19+ B cell		
Sfpi1	2.09	Essential regulator of B-cell differentiation [175]
Pou2f2	2.08	Required for T-cell independent B cell activation [33]
Spic	1.98	Promotes B cell differentiation [161]
Pou2f3	1.94	Unknown function in B cell, but has almost the same binding preference as Pou2f2
Elf4	1.70	Regulates proliferation of B cells [85]

used at all in our analysis.

Two broad strategies have been developed to extract TF binding properties from PBM data. The first strategy involves estimation of PWM using the PBM data. Berger et al. [17] developed a rank-based statistic called E-score to quantify a TF binding strength to 8-mers, and a the constructed a PWM from the 8-mers with the highest E-scores. Zhao et al. [204] developed BEEML-PBM, a biophysical model of TF-probe binding, that directly estimates the binding energy from the PBM data while accounting for several experimental artifacts. The second strategy does not fit a single PWM; instead it uses information of all k-mers from the PBM data to predict TF binding to any sequences. The simplest method from this category sums over the strength (defined as median probe intensities) of all k-mers in a sequence [207]. More sophisticated methods have also been proposed. For example, Agius et al. [1] developed a support vector regression (SVR)-based method using a novel string kernel. A recent large scale comparison of dozens of methods for using PBM data has identified BEEML-PBM as one of the best methods in both reproducing *in vitro* and predicting *in vivo* binding.

There has been considerable debate on whether the binding specificity of a TF can be represented by a single PWM. While Badis et al. [6] reported that about half of the TFs in mouse have two distinct sets of binding profiles, Zhao et al. [204] contended that once the experimental artifacts were removed, most of the variation in the PBM data could be explained by predicted binding affinities based on a single PWM. We believe the best metric for addressing this question is the predictive accuracy on *in vivo* TF binding from ChIP-seq data. Using this metric, we show that the PLAR-PBM generally outperform methods imposing a single PWM. The advantage of PLAR-PBM over other k-mer based ones (Figure 3.3) results from its several features: the sparse linear model that helps avoid overfitting, the biophysical nature of the model that allows the experimental artifacts (e.g. biases due to position in the probe) to be easily addressed. In this way, PLAR-PBM combines the strengths of both the biophysical approach (BEEML-PBM) and the k-mer based methods.

We demonstrated that incorporating additional information, most notably, chromatin accessibility from DNase I HS data, further improves the accuracy of TF binding prediction. While the utility of the DNase data has been demonstrated before (e.g. [76]), the simple method of intersecting the DNase I hypersensitive sites and the PWM matches tends to lose a significant amount of information as we demonstrated in Figure 3.5. In contrast, our model uses a probabilistic framework which improves the accuracy of the predictions. A somewhat unexpected observation from our experiments is that evolutionary conservation did not provide additional benefits (beyond the ones obtained from using the DNase data) for TFBS predictions. One explanation is that even functional TFBS may undergo rapid turnover over evolutionary timescale [155].

Several recent papers explored related ideas. Chromia [193] used a hidden Markov model to combine sequence-specific TF binding with histone modification data, but their predictions were based on PWM scoring and only focused on a dozen of TFs in mouse embryonic stem cells. CENTIPEDE [139] used a graphical model to integrate TF-DNA interaction, epigenetic and evolutionary data. However, CENTIPEDE also used PWM representation, and their predictions were focused only on lymphoblast cell lines. Similarly, Ernst et al. [48] combined experimental data from a number of tissues to generate a single (global) TF-target prediction map. However, that method has also relied on PWMs and no tissue specific predictions were made. Another recent strategy of analyzing PWM and DNase data relies on a "footprint" in the data that TFs leaves and that could be experimentally detected [125]. Recently, Neph et al. [124] used this data and PWMs to predict TF-TF interactions (though not TF-gene interactions) across a large number of human tissues. However, DNase I footprint data is only available for few tissue types, and similar strategy to overlap PWM hits with DNase data does not perform as well, as we demonstrated in Results. It is not entirely clear that such footprint is a universal feature of all TFs and to identify such footprints in mouse would require sequencing data with very high coverage, which is not currently available for most of the tissues we analyzed.

Our work represents the first major effort to provide a systematic map of all targets for hundreds of TFs across a large number of tissues. We complied a resource that provides TF-target predictions for all 284 TFs studied across the 55 tissue/cell types (Section 3.6.6) and Supplementary Website). We hope that such comprehensive resource would prove useful for researcher studying specific tissues or attempting to reconstruct regulatory networks in such tissues using one of several network modeling methods rely on TF-gene interactions to seed their models [46, 94, 107, 158].

Chapter 4

DREM 2.0 and analysis of dynamic gene responses in arabidopsis following ethylene treatment

Transcription factors (TFs) regulate gene expression by binding to their promoters, and together the TFs and their target genes form a complicated transcriptional regulatory network (TRN). In the previous chapter, we developed methods to computationally predict tissue-specific TRNs. However, in eukaryotes, such networks are not only tissue-specific, but also dynamic with different (overlapping) sets of transcription factors activating genes at different points in time or developmental stages. Reconstructing such dynamic networks requires the integration of different types of data and is a non-trivial task. In this chapter, we extend previous works by our group on DREM for inferring dynamic TRNs [46] (see below) by integrating it with DECOD that was developed in Chapter 2 to allow discriminative motif finding within DREM, and also allowing the use of time course binding data. In addition, the tissue-specific TRNs that were made available by methods presented in the previous chapter can also be used as input to model tissue-specific TRNs with expression data in the corresponding tissue. We apply the new method to analyze dynamic gene responses in arabidopsis following ethylene treatment.

4.1 Introduction

4.1.1 Extending DREM

Many methods have been proposed for reconstructing TRNs ([51, 89, 108]). Most of such methods involve the integration of static gene expression measurements with TF-DNA binding data either from computational predictions or from existing databases, with a focus on reconstructing static TRNs that does not change with time. Such static TRNs are not realistic in nature, particularly in what happens following treatments with stimulants or in processes that are temporal themselves such as cell cycle and development.

Recently, more and more studies measure gene expression changes over time (time series gene expression, [10]) in many species using either microarray [56, 154] or RNA-seq [116, 178]. Several methods have been developed specifically for analyzing such time-series expression data from various perspectives including unsupervised clustering [34, 45, 171], detecting differentially expressed genes [11, 31, 90] and reconstructing dynamic TRNs by using the time series expression data alone [9, 96, 176, 190] (see [12] for a comprehensive review). However, compared with the availability of such time series expressions data, most TF-DNA interaction data are still static. Combining the dynamic expression data with the static TF-DNA interaction data remains a computational challenge.

To provide a general method that can be widely applied to reconstructing dynamic regulatory networks, our group previously developed DREM [46] (*d*ynamic *r*egulatory *e*vent *m*iner), a method that integrates times series and static data using an Input-Output Hidden Markov Model (IOHMM) [15]. DREM learns a dynamic TRN by identifying bifurcation points, places in the time series where a group of co-expressed genes begins to diverge. These points are annotated with the TFs controlling the split, leading to a combined dynamic model that can determine when TFs activate genes and what genes they regulate. Since its release 5 years ago, the DREM software has been used for modeling a wide range of GRNs for example stress response in yeast [46] and E. coli [47], development in fly by the modENCODE consortium[116], stem cell differentiation in mice [114] and disease progression in human [61].

While DREM has been successfully used for multiple species, there are several limitations in the original version of DREM. As mentioned above, DREM identifies bifurcation points at which the expression profiles of two sets of coexpressed genes begin to diverge, and DREM then associates TFs with such splits. But for some splits, no TFs would be annotated. This may happen when the TF-DNA interaction data is incomplete and so is unable to explain a split node that DREM identifies. In the original DREM, no further investigation could be performed for such occasions. It would be useful to be able to apply discriminative motif finding tools like DECOD [70] (Chapter 2) on such split paths and search for *de novo* motifs that are present in one path but not in the other, and match such motifs with potential TF binding sites. Moreover, the original DREM only supports the use of static TF-DNA binding data. Although most TF-DNA interaction data is still static, dynamic binding data measured by ChIP experiments is becoming available [127, 191]. The ability to incorporate such dynamic binding data would lead to more accurate dynamic TRN models.

To address these issues, we developed a new version of DREM, DREM 2.0, that enables running DECOD for discriminative motif finding within DREM and also enables the use of dynamic TF-DNA binding data. These, together with other features including the availability of more comprehensive TF-DNA binding data for many species, the ability to use the expression levels of TFs in the modeling, and the ability to use continuous (instead of binary) binding data, are incorporated into DREM 2.0 [158].

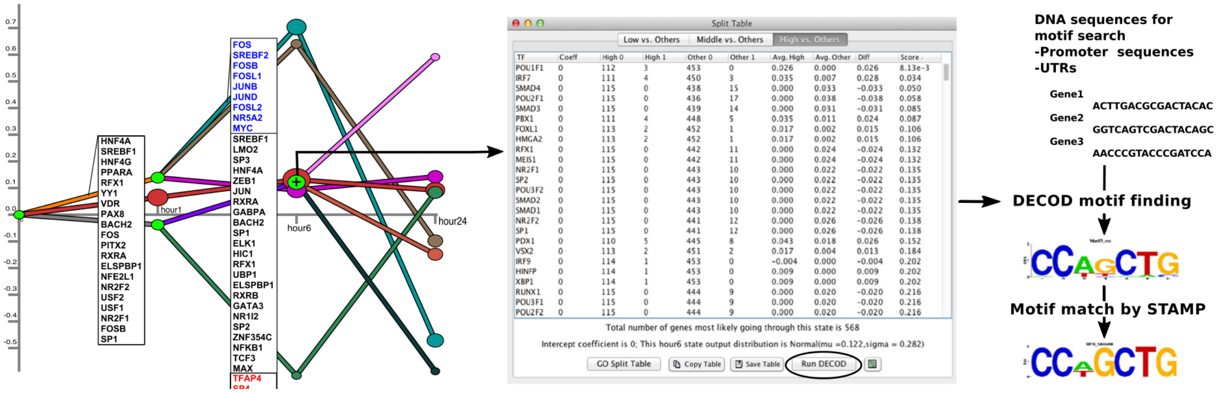
4.1.2 Gene responses in arabidopsis following ethylene treatment

In plant, a simple hydrocarbon gas, ethylene, regulates many biological processes including fruit ripening, stem cell division, differential cell growth, stress and pathogen responses, and senescence, etc [79]. Despite its importance, we lack a comprehensive understanding of how ethylene mediates this myriad of morphological responses. The dynamic nature of the ethylene response, a rapid growth inhibition independent of the master transcriptional regulator ETHYLENE INSENSITIVE3 (EIN3), followed by an EIN3-dependent sustained growth inhibition, calls for a temporal study of the ethylene response [18]. The transcription factor EIN3 is necessary and sufficient for the ethylene response and accumulates upon a duration of exogenous ethylene gas treatment [62]. Although hundreds of ethylene response genes have been identified, because some targets of EIN3 are transcription factors (e.g. ETHYLENE RESPONSE FACTOR1 (ERF1)), it is challenging to distinguish immediate early targets from those further downstream. To understand the dynamics of the EIN3-mediated ethylene transcriptional response, we performed a genome-wide study of dynamic ethylene-induced EIN3 protein-DNA interactions using chromatin immunoprecipitation followed by sequencing (ChIP-Seq) and simultaneously determined the repertoire of target genes that are transcriptionally regulated by ethylene (RNA-Seq) in *Arabidopsis thaliana*. Tracing the transcriptional cascade, we used DREM 2.0 which allows the use of dynamic binding data to investigate if EIN3-mediated genes contribute to a component of the ethylene transcriptional response.

4.2 Method: DREM 2.0

4.2.1 Integrating DECOD with DREM

During learning DREM assigns genes to paths in the network model and uses split nodes to represent sets of genes that change their expression between consecutive time points. TFs are assigned to split nodes allowing DREM to infer their time of activation. When the protein-DNA interaction data is unable to explain some of the split nodes (i.e. no TF is assigned to that split), it could mean that the interaction data is incomplete. To still allow the identification of such TFs, we integrated with DREM 2.0 the discriminative motif finder DECOD [70] (Chapter 2). Now user can search for discriminative DNA motifs between promoters of genes assigned to diverging paths emerging out of any split node, and the motifs that DECOD reports can be matched against known motif databases directly within DECOD in a user-friendly manner using STAMP [106] (Figure 4.1).



TF	Coeff	High 0	High 1	Other 0	Other 1	Avg. High	Avg. Other	Diff	Score
POU1F1	0	112	3	453	0	0.026	0.000	0.026	8.13e-3
IRF7	0	111	4	450	3	0.035	0.007	0.028	0.034
SMAD4	0	115	0	438	15	0.000	0.033	-0.033	0.050
POU2F1	0	115	0	436	17	0.000	0.038	-0.038	0.058
SMAD3	0	115	0	439	14	0.000	0.031	-0.031	0.085
PBX1	0	111	4	448	5	0.035	0.011	0.024	0.087
FOXL1	0	113	2	452	1	0.017	0.002	0.015	0.106
HMGA2	0	113	2	452	1	0.017	0.002	0.015	0.106
RFX1	0	115	0	442	11	0.000	0.024	-0.024	0.132
MEIS1	0	115	0	442	11	0.000	0.024	-0.024	0.132
NR2F1	0	115	0	443	10	0.000	0.022	-0.022	0.135
SP2	0	115	0	443	10	0.000	0.022	-0.022	0.135
POU3F2	0	115	0	443	10	0.000	0.022	-0.022	0.135
SMAD2	0	115	0	443	10	0.000	0.022	-0.022	0.135
SMAD1	0	115	0	443	10	0.000	0.022	-0.022	0.135
NR2F2	0	115	0	441	12	0.000	0.026	-0.026	0.138
SP1	0	115	0	441	12	0.000	0.026	-0.026	0.138
PDX1	0	110	5	445	8	0.043	0.018	0.026	0.152
VSX2	0	113	2	451	2	0.017	0.004	0.013	0.184
IRF9	0	114	1	453	0	-0.004	0.000	-0.004	0.202
HINFP	0	114	1	453	0	0.009	0.000	0.009	0.202
XBP1	0	114	1	453	0	0.009	0.000	0.009	0.202
RUNX1	0	115	0	444	9	0.000	0.020	-0.020	0.216
POU3F1	0	115	0	444	9	0.000	0.020	-0.020	0.216
POU2F2	0	115	0	444	9	0.000	0.020	-0.020	0.216

Total number of genes most likely going through this state is 568

Intercept coefficient is 0; This hour6 state output distribution is Normal(mu =0.122,sigma = 0.282)

Gene1 ACTTGACGCGACTACAC
Gene2 GGTCAGTCGACTACAGC
Gene3 AACCCGTACCCGATCCA

Figure 4.1: DECOD motif search in DREM 2.0. (left) DECOD motif search was performed for one node (+ sign). (middle) After clicking the node, the DREM split table opens which shows the enrichment of TFs on gene sets divided by the split. As this split has three outgoing paths, DECOD can be run in three different ways. Here, we compared genes in the highest path against the other two paths (Tab High vs. Others) by clicking the Run DECOD button (circled). (right) Illustration of one of the TF motifs found by DECOD and its most similar match in TRANSFAC according to STAMP

4.2.2 Allowing the use of dynamic TF-DNA binding data

In the original DREM, only static TF-DNA interaction tables are supported (Figure 4.2a). The underlying Input-Output Hidden Markov Model learning can now accommodate dynamic TF-DNA binding data for each time point in DREM 2.0 in the following way. The transition probabilities for the IOHMM are derived from a logistic regression classifier that uses the protein-DNA interaction data as supervised input and utilizes them to classify genes into diverging paths at a split node in the model. In the new version the nodes in the input layer can be dynamic, so that at different times, a different set of TF-DNA binding data can be used, and the logistic function can depend on input from the specific time point it is associated with (Figure 4.2b). Since dynamic binding data is often only available for a (small) subset of TFs, DREM 2.0 supports a joint static-dynamic input format for TF-DNA interactions as well (Figure 4.2c).

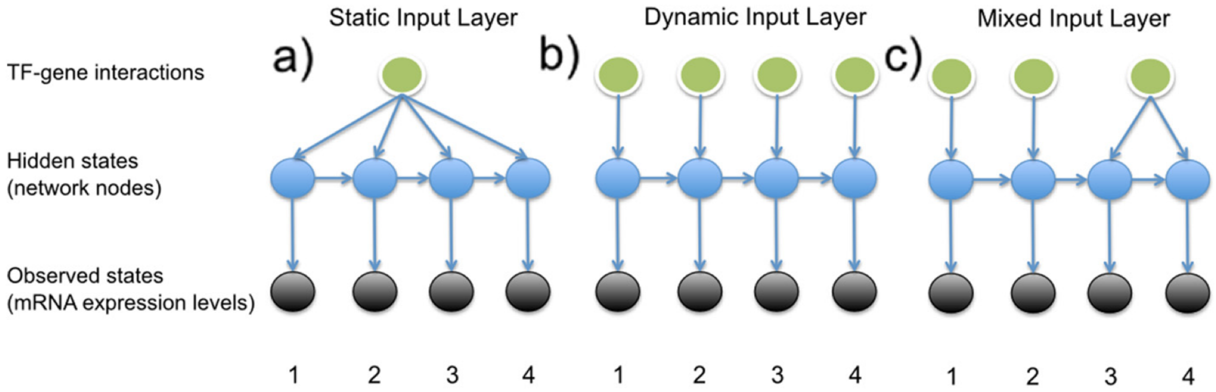


Figure 4.2: Possible IOHMM topologies in DREM 2.0. The basic topology for a DREM 2.0 IOHMM is shown. The hidden states represent the network nodes (in blue) that we are interested in. The observations (black nodes) are the gene expression ratios which are given to the model, these are dynamic and dependent on the time point. The protein-DNA interaction data (green nodes) are used as supervised input data to construct the network. (a) In the original DREM formulation only one static input node is connected to all hidden nodes. In DREM 2.0 the nodes in the input layer can be dynamic and dependent on the time point with a topology either fully dynamic (b) or a mix of static and dynamic input.

4.3 Method: ethylene responses in arabidopsis

The ChIP-seq and RNA-seq experiments (from 4.3.1 to 4.3.8) were performed and analyzed by collaborators in Salk Institute, and relevant details are provided below for completeness.

4.3.1 Plant material

The Arabidopsis thaliana ecotype Columbia (Col-0) was the parent strain for these experiments. Genotypes used for this study include wildtype Col-0, and mutants *ein3-1* [25], *ein3-1/eil1-1* [4], *hls1-1* (*hls1*) [91], *hlh1*, *hlh2*, *hlh3*.

4.3.2 Growth of Arabidopsis seedlings

Three-day-old etiolated seedling tissue was used for these experiments unless otherwise noted. Seeds were sterilized and sown on Murashige and Skoog (cat#LSP03, Caisson) media pH 5.7, containing 1% sucrose and 1.8% agar. After stratification for three days in the dark at 4°C, exposure to light for 2-4 hours to induce germination, seeds were dark-grown in hydrocarbon free air at 24°C for three days. Etiolated seedlings were subsequently treated with ethylene gas at 10 μL L^{-1} for 0, 0.25, 0.5, 1, 4, 12, and 24 hours.

4.3.3 Chromatin preparation and immunoprecipitation

Etiolated seedlings were collected in the dark, immersed in 1% formaldehyde solution, and cross-linked under vacuum for 15 minutes. A final concentration of 125 mM glycine was used to quench the formaldehyde for 5 minutes under vacuum. Cross-linking under vacuum resulted in translucent etiolated seedling tissue. Tissue was liquid nitrogen ground and extraction of chromatin was performed as described in [99].

Chromatin immunoprecipitation (ChIP) was performed as described in [99] with modifications, including the use of the Bioruptor sonicator (Diagenode). Bioruptor settings used were: H, 25 cycles of 0.5 min on, 0.5 min off, with 5 minute rests between every 5 cycles. Sonication was performed in a cooling water bath at 4°C. A small amount of chromatin (10 μl) was evaluated for shearing; the size range of chromatin was 150-700 bp, the majority of fragments at 300-400 bp.

Affinity-purified rabbit polyclonal antibodies capable of detecting the C-terminus of EIN3 were used in immunoprecipitation reactions. Details regarding the generation of EIN3 antibodies were previously described [62]. Prior to the experiments in this study, the amount of purified EIN3 antisera per immunoprecipitation reaction was optimized and 8 μl of purified EIN3 antisera was determined to yield the optimal enrichment of the ERF1 promoter, the known target of EIN3 (data not shown). We then substituted Dynabeads Protein A (Invitrogen, cat#100-1D) and Dynabeads M-280 Sheep anti-Rabbit IgG (Invitrogen, cat#112-04D) for the salmon sperm DNA blocked Protein A agarose beads recommended in the protocol [99], as to avoid sequencing of salmon sperm DNA. Immunoprecipitation and washing of Dynabeads were performed using the buffers in [99], otherwise Dynabeads were used as per the manufacturers instructions. Multiple pipetting steps were performed while washing the beads to reduce non-specific binding carryover. Resulting ChIP DNA was purified as in [99].

Quantitative PCR revealed that relative ChIP enrichment for the promoter of ERF1 performed with the Dynabeads M-280 Sheep anti-Rabbit IgG was higher in comparison to Dynabeads Protein A. Thus, Dynabeads M-280 Sheep anti-Rabbit IgG was used in all subsequent experiments. Primers for the ERF1 promoter encompassing the EIN3 binding site, are as follows: F-GGGGGCATGTATCTTGAATC, R-TGCTGGATCAACTCAACAAAA. Actin primers were as in [110]. Enrichment was calculated using the Delta-Delta-Ct method with normalization to the reference Actin; fold change was calculated relative to the control for non-specific binding (EIN3 ChIP performed in ein3-1 mutant).

ChIP was performed in chromatin derived from wildtype Col-0 three-day-old etiolated seedlings treated with 0, 0.25, 0.5, 1, 4, 12, and 24 hours of ethylene. Two independent biological replicates were used in two replicates experiments for timepoints, 0, 0.5, 1, 4 hours ethylene gas treatment. Single replicates exist for 0.25, 12, 24 hours of ethylene gas treatment.

4.3.4 Total RNA extraction

Total RNA was extracted from liquid nitrogen ground etiolated seedlings using the Qiagen RNeasy Plant Mini Kit with Qiashredder columns (cat#74904), with DNaseI (Qiagen, cat#79254) treatment prior to RNA precipitation in sodium acetate and ethanol. Concentrations of RNA were determined using the ND-1000 spectrometer (Nanodrop). Experiments were performed in three biological replicates for timepoints, 0, 0.25, 0.5, 1, 4, 12, 24 hours ethylene gas treatment.

4.3.5 ChIP-Seq library generation and sequencing

Resulting ChIP DNA from two pooled ChIP reactions above was used to generate a sequencing library as per the Illumina ChIP-Seq manufacturers instructions. The Illumina Genome Analyzer II was used to sequence the single-read ChIP-Seq libraries as per manufacturers instructions, for 36-43 bps. Raw sequencing data was analyzed using the Genome Analyzer Pipeline v.1.4.0.

4.3.6 PolyA selection and mRNA-Seq library generation

At least 80 μg total RNA was subject to polyA selection using the Poly(A)Purist MAG Kit (Ambion, cat#AM1922). PolyA RNA was subsequently concentrated by ammonium acetate ethanol precipitation and concentrations were determined using the Qubit fluorometer (Invitrogen) and the Quant-iT RNA Assay Kit (Invitrogen, cat#Q33140). 50-100 ng of polyA RNA was used in a strand-specific library preparation as per the SOLiD Total RNA-Seq Kit protocol (Invitrogen, cat#4445374) and AMPure XP beads (Agencourt, cat#A63881) were used for purification of cDNA and amplified DNA. Samples were barcoded for multiplexing using the SOLiD RNA Barcoding Kit (Invitrogen, Module 1-16 cat#4427046, Module 17-32 cat#4453189, Module 33-48 cat#4453191) as per manufacturers instructions; final size selection was performed using AMPure XP beads instead of the PAGE purification recommended in the protocol. Size selected libraries were then purified using the MinElute Gel Extraction Kit (Qiagen, cat#28604). Resulting concentrations of libraries were detecting using the Qubit fluorometer and Quant-iT dsDNA

High-Sensitivity Assay Kit (Invitrogen, cat #Q33120). RNA libraries were sequenced for 50 bps on the SOLiD4 platform (Life Technologies).

4.3.7 ChIP-seq data analysis

The Illumina GERALD module was used to align the sequenced reads to the Col-0 reference genome, version TAIR10 (ftp://ftp.arabidopsis.org/). The analysis variable for the ELAND alignment program was set to eland_extended, as read length was greater than 32 bases (e.g. 36-43). Resulting aligned unique single copy reads were used in ChIP-Seq peak analysis .

Saturation analysis of the ChIP libraries conducted using the spp software [80] revealed that all samples were at least within 15% of saturation. Peak analysis was performed individually on each timepoint in each biological replicate using the corresponding 0 hour ethylene treated wildtype Col-0 EIN3 ChIP sample as a control. Two additional ethylene treated (4 hour) wildtype EIN3 ChIP biological replicates were included in the analysis, with corresponding mutant *ein3-1* ethylene treated (4 hour) EIN3 ChIP samples as controls. Three software packages: spp [80], MACS [203], PeakSeq [152] were originally used to identify peaks/regions of binding. Parameters for each software were as follows: MACS (p-value = 0.01), spp (FDR = 0.1), PeakSeq (FDR = 0.1, mingap = 200, minhit = 20, minratio = 3.5). Binding regions were merged when the maximum gap between two peaks was less than 200 bp determined by separate software packages. Subsequent analysis was performed in R. Overlapping peaks in one biological replicate in one timepoint by more than one software package were retained as binding regions. Because of the variation of the number of called peaks in each software and each timepoint, we used a majority vote to call peaks to identify all high stringent EIN3 targets. PeakSeq results differed significantly from spp and MACS (12-76%), therefore only spp and MACS were ultimately used.

Using this method, 1460 EIN3 binding regions were identified. For each EIN3 binding region, the reads per kbp of binding site per million sample reads (RPKM) were calculated. Median normalization of the RPKM values between timecourse biological replicates was performed in R.

Resulting RPKMs were log2 transformed with respect to the 0 hour ethylene treatment wildtype Col-0 EIN3 ChIP. Normalization with respect to an input genomic control did not produce distinctively different EIN3 binding pattern profiles (data not shown). EIN3 binding regions were then associated to a gene if located within 5 kbp. The nearest expressed gene (RPKM $\geq$ 1) was assigned if there were more than one gene within 5 kbp. If both genes were not expressed, the nearest gene was selected. Distance was determined from the binding region center to the gene feature using the TAIR10 annotation (ftp://ftp.arabidopsis.org).

4.3.8 mRNA-Seq analysis

The SOLiD Bioscope v1.3 software was used to align the reads to the Col-0 reference genome TAIR10 (ftp://ftp.arabidopsis.org/). Two perfect matches per location were allowed. Exonic expression was determined (RPKM) using mRNA-Seq reads mapping in exons in the direction of transcription. Genes were denoted as expressed if they contained RPKM values greater than one for at least one biological replicate in one timepoint. Differentially expressed genes were then called (t-test p-value = 0.05, 50% difference from prior timepoint of ethylene gas treatment), and log2 normalized with respect to the 0 hour ethylene gas treatment control.

4.3.9 Data analysis with DREM 2.0

In order to reconstruct the dynamic regulatory networks that were activated following ethylene treatment, we used DREM 2.0 to analyze the ethylene transcriptional responses as measured by RNA-seq using a combination of the dynamic EIN3 binding data and static TF-DNA binding data for other arabidopsis TFs. To obtain the static interaction data, we extracted 11,355 TF-DNA interactions from the AtRegNet AGRIS database [198]. For each EIN3 target gene, the average RPKM values from two input control samples at 0 and 4h were used as a cutoff to determine whether it was bound by EIN3 or not at each time point. We ran DREM 2.0 using the RNA-seq data allowing for 3-way splits. We filtered out genes that did not change at least 2-fold (up or down) at any time point, and we used the default values for all other parameters.

4.3.10 Hormone-related genes

To identify genes associated with a hormone signal or response (hormone-related), we used the annotation in the Arabidopsis Hormone Database [138](http://ahd.cbi.pku.edu.cn/) in addition to other datasets including relevant ethylene microrarray studies in etiolated seedlings [3, 123]. The amount of genes involved in hormone responses in the genome was 21% (5729/27416), where as the amount of genes involved in our EIN3 target group was 46%.

4.4 Result

4.4.1 ChIP-seq of EIN3 binding and RNA-seq of gene responses following ethylene treatment

We performed ChIP-Seq using a native antibody that recognizes EIN3 [62] and mRNA-Seq in three-day-old dark grown seedlings during a timecourse of ethylene treatment (0, 0.25, 0.5, 1, 4, 12 and 24 hours following the treatment). We identified 1460 EIN3 binding regions in the Arabidopsis genome associated with 1314 genes by stringent analysis of the temporal ChIP-Seq data (Section 4.3.5 and 4.3.7). Genes associated with EIN3 binding regions are referred to as EIN3 targets hereafter.

Overall, of the 1314 EIN3 target genes, 29% (381) were transcriptionally ethylene-regulated (EIN3-R; t-test p-value $\leq$0.05 and fold difference $\geq$50% compared with time 0h), 67% (880) were not transcriptionally ethylene-regulated (EIN3-NR) and a negligible amount of target transcripts were below detection (EIN3-ND, 4%). This is consistent with previous reports that transcription factor binding does not necessarily coincide with changes in transcription [103], especially for master regulators targeting other transcription factors or other factors involved in chromatin state regulation. The majority of studies that exist in the literature have shown that

EIN3 acts as an activator. We observed this activation at the genome-wide level (Figure 4.3). We found that a majority of EIN3-R are induced (85%). We observe over-representation of up-regulation of EIN3 targets when compared to the regulation of all genes that respond to ethylene (Figure 4.3 and 4.4A). Many EIN3-R are transcription factors ($\sim$14%), and EIN3 targets are significantly enriched in gene ontology (GO) terms related to transcription factor regulation, confirming that EIN3 initiates a transcriptional cascade (Figure 4.4B) [25]

4.4.2 DREM 2.0 identifies waves of activated transcriptional regulation regulated by EIN3

We used DREM 2.0 to analyze this time course transcriptional ethylene response data, taking advantage of its ability to integrate the dynamic EIN3 binding data with other static TF-binding data extracted from existing databases (see Section 4.3.9 for details). The average expression levels for each trajectory in the model that DREM identifies are shown in Figure 4.5. DREM identifies that the ethylene response occurs in four waves of transcription significantly regulated by EIN3 (overall pathway hypergeometric p-value $< 10^{-10}$, the top four paths in Figure 4.5 and also shown in Figure 4.6A). These waves display distinct temporal transcription behaviors, and reduction of transcriptional noise occurs in successive temporal waves (Figure 4.6B and Figure 4.7). Genes were enriched in specific biological functions within these four transcriptional waves (hypergenometric p-value $< 10^{-3}$), e.g. RNA binding/translation (Wave 1, Wave 3), cell wall maintenance (Wave 2), response to endogenous stimulus (Wave 4). The second wave is enriched for genes involved in cell wall maintenance, and the expression of these genes steadily increases following one hour of ethylene treatment, consistent with kinetics of EIN3-dependent growth inhibition [18].

The four waves of the ethylene transcriptional response each contain a unique subset of EIN3 targets. The first wave is highly variable, lower in steady-state levels of transcription, and also contains the lowest percentage of EIN3 targets and hormone-related genes (Figure 4.6B). The

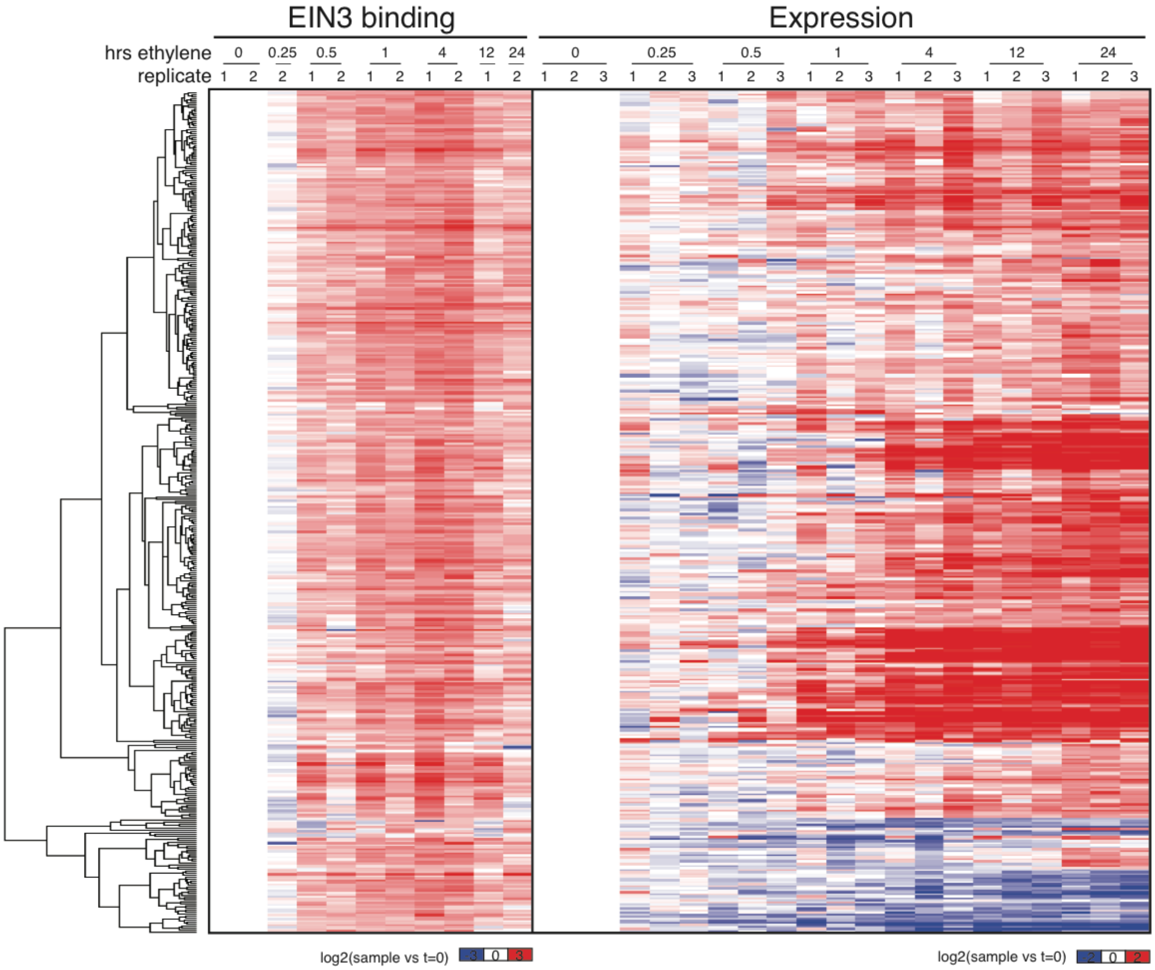


Figure 4.3: Patterns of EIN3 binding and expression of ethylene-regulated targets are strikingly evident over a timecourse of ethylene gas treatment. EIN3 binding increases with ethylene treatment to a maximum at 4 hours of ethylene treatment for all targets. Each line in the heatmap represents the logarithm RPKM value for the representative EIN3 binding site (left panel) and transcript (right panel).

next three waves of transcription are successively less variable and contain higher percentages of EIN3 targets and hormone-related genes. The four waves of ethylene-induced transcription account for 50% of the transcriptionally ethylene-regulated EIN3 targets (EIN3-R), the remaining EIN3 targets are distributed among other patterns of transcription that do not contain significant numbers of EIN3 targets in each transcriptional trajectory (Figure 4.5). The expression kinet-

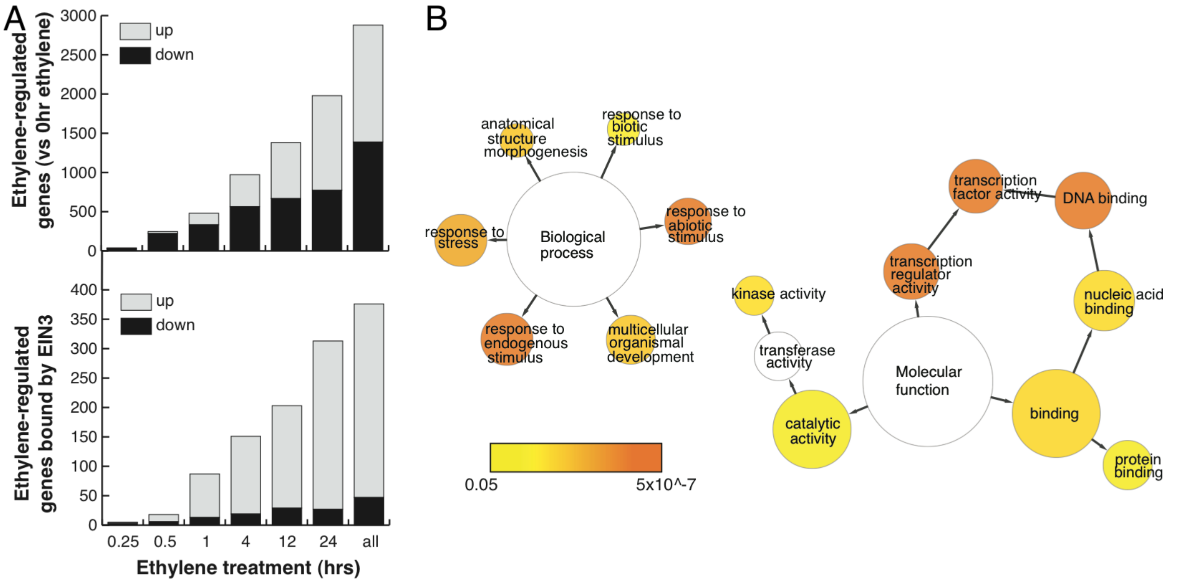


Figure 4.4: (A) The number of ethylene-regulated genes. (Upper panel) Equivalent numbers of genes are up- and down-regulated upon ethylene treatment. (Lower panel) Majority of EIN3 targets differentially expressed upon ethylene treatment are up-regulated. (B) Functional categories over-represented for EIN3 targets that are ethylene-regulated (EIN3-R). Network was generated using BiNGO ([104], v 2.44) using the GOSlim_Plants ontology, Benjamini and Hochberg p-value legend is indicated below.

ics and reduction of transcriptional noise we observe in the ethylene-induced waves can be tied to distinct mechanisms of transcriptional control, or may reflect heterogeneity of the ethylene response in different tissues, which can be resolved using single cell analysis. Taken together, it appears that the initial early ethylene transcriptional response is noisy and less focused functionally, whereas during exogenous ethylene application, EIN3 accumulates, and the established ethylene transcriptional response is hormone-focused and less noisy.

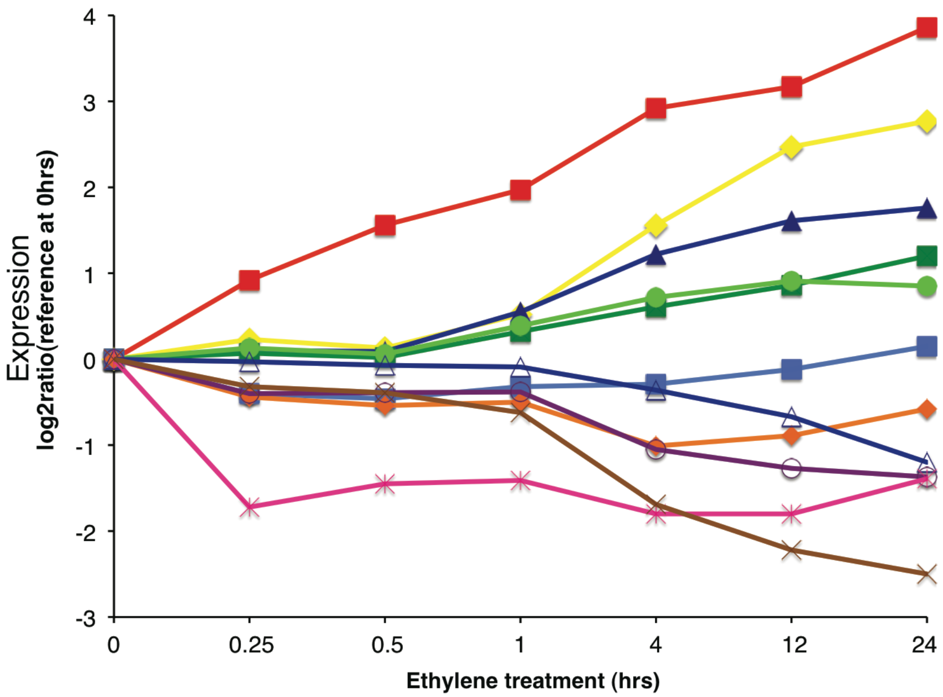


Figure 4.5: Ethylene-regulated transcription kinetics from DREM analysis. EIN3-modulation is significant (10^{-10}) in the four trajectories with the highest expression ratios, indicated by red, yellow, dark blue, dark green lines.

4.5 Discussion

Up to now, most studies that investigate TF binding have been static and few dynamic binding studies have been conducted[127, 145, 208]. Integrating such dynamic binding data with time course expression data to elucidate the underlying transcriptional regulatory network has been a major computational challenge. By extending DREM [46], we present the first tool, DREM 2.0 [158], that reconstructs the dynamic regulatory networks using such data. The ability to incorporate temporal binding data allows DREM to reduce false positive assignments by only assigning TFs that are active at that time point (based on the time points binding data). This in turn can both help identify co-regulators for which only computational predictions exists and also lead to the identification of different waves of transcriptional regulation, where the same

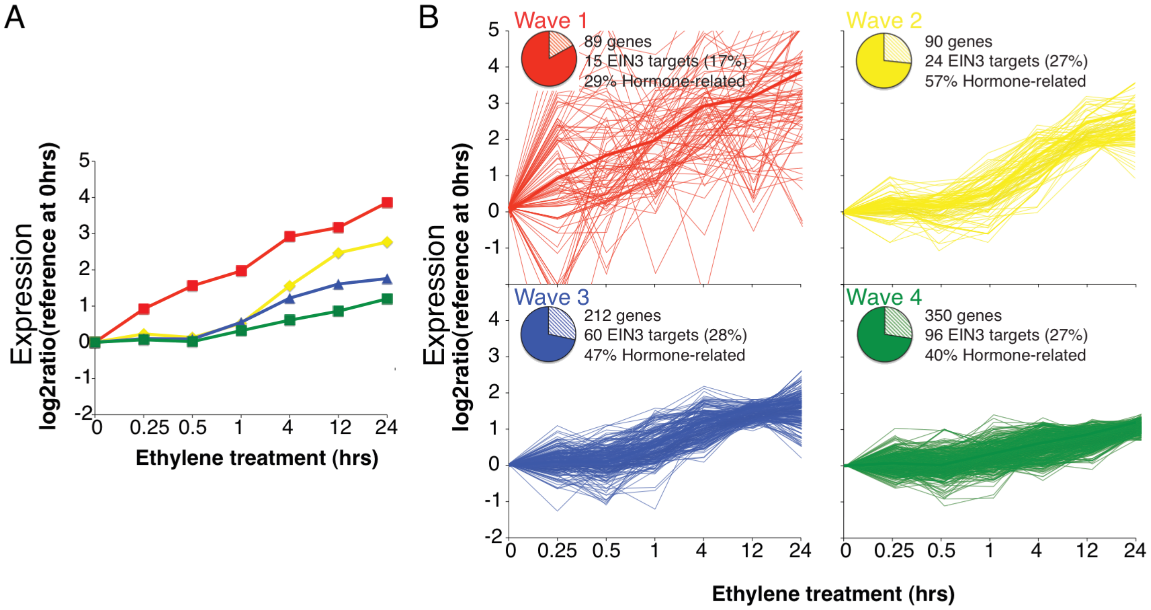


Figure 4.6:) DREM paths representing waves of induction of steady-state levels of transcription by ethylene for genes that are regulated by EIN3, implicating different modes of transcriptional regulation in the ethylene response. Right panels contain all genes for each wave.

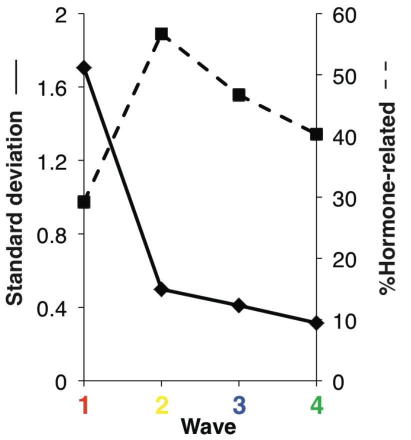


Figure 4.7: The EIN3-modulated ethylene transcriptional response occurs in four waves with various levels of noise. A decrease in standard deviation correlates to an increase of hormone-related genes.

TFs activate different sets of genes at different time points. Using DREM 2.0, we found that in arabidopsis, transcription factor binding upon a timecourse of ethylene treatment resulted in an induction of EIN3 binding for all targets. The ethylene transcriptional response occurred in waves of transcription that were temporally distinct, variable in the amount of noise, and significantly regulated by EIN3.

In addition to allowing dynamic binding data, DREM 2.0 also incorporates the *de novo* discriminative motif finding tools DECOD (Chapter 2, [70]) to allow the discovery of previously undocumented transcription factors that drives the expression divergence of genes at specific time points. Other new features of DREM 2.0 includes the availability of binding data for more species and the support of using continuous instead of binary binding data [158]. Like its predecessor, DREM 2.0 also comes with an easy-to-use GUI. Together, these new features and improvements will make DREM 2.0 a more widely used software package for studying dynamic regulatory networks.

Chapter 5

Conclusions and future works

5.1 Summary of contributions

The flow of information from DNA to proteins is regulated at multiple levels. Transcriptional regulation directly controls the rate at which RNA molecules are synthesized from DNA templates, and it is one of the first and key steps in regulating such information flow. In this thesis, we have presented several computational methods and analyses that facilitate a better understanding of transcriptional regulation from different perspectives.

We first developed DECOD, a fast and accurate method for discriminative motif finding. DECOD can be used to study differential regulation, where two groups of genes show different transcriptional behaviors under the control of different TF or cofactors. This happens when two sets of coexpressed genes start to diverge in their expression profiles after a particular time point in time series expression analysis, and when two sets of genes show similar expression profiles in one condition but distinct profiles in another, etc. DECOD can also be used to discover binding motifs for TFs with unknown binding preference from large scale sequencing data generated by experiments like ChIP-seq when a proper set of background sequences is used. The ability to scale up well to such large number of input sequences is one of the key features that distinguishes DECOD from many other tools for similar purposes. Using DECOD on a new experimental dataset that studies p53 mutant binding, we were able to identify several p53 co-factors and suggest new mechanisms about p53 activation. DECOD has a GUI interface that can be run across different platforms, and can be easily used by experimental biologists with little computational experience. In addition, DECOD is linked with the motif matching online tool STAMP [106], so that the user can directly match a discriminative motif discovered to known TFBS databases within DECOD. These features make DECOD a valuable computational tool to study TF binding and transcriptional regulation.

We next developed a new method, PLAR-PBM, which uses protein binding microarray data to infer TF binding profiles. PLAR-PBM uses a biophysically motivated model to predict the binding probabilities to individual short k-mers from the PBM data for a TF. PLAR-PBM does not depend on a PWM model which assumes independence between positions, and it outperforms several other methods when classifying known binding sites from ChIP-seq data. We then used an integrated model that combines DNase I hypersensitivity data and also possibly conservation data with PLAR-PBM to predict *in vivo* tissue-specific TF activities and binding sites. We were able to accurately predict top TFs in several tissues including liver, retina, and B cells, etc. that have known functions specifically in these tissues. Up to now, most transcriptional regulatory network studies have used only general binding data, and our work is one of the first to computationally predict tissue-specific TF binding sites. The ability to quantify and predict tissue-specific TF activities further enables the discovery of potentially novel biological functions for the top TFs in each tissue with unknown functions. We provide a comprehensive resource for computationally predicted tissue-specific binding sites for hundreds of mouse TFs across 55 tissues/cell lines. This will be of value to the study of tissue-specific transcriptional regulatory networks.

Finally, we presented DREM 2.0 that combines time-series expression data with binding data to analyze dynamic regulatory networks. Compared with its predecessor, DREM 2.0 integrates with DECOD to allow discriminative motif finding directly on two sets of genes showing diverged expression patterns after a specific time point but with no known TF regulations from the binding data, and it is also able to use dynamic and continuous binding data to more accurately model the transcriptional regulatory network and assign TFs to specific time points based on when they are activated. Using DREM 2.0, we analyzed an experimental dataset measuring gene responses in arabidopsis following ethylene treatment with dynamic EIN3 binding data, and identified interesting patterns of transcriptional activation regulated by EIN3 in successive waves. DREM 2.0 comes with binding data for several more species collected from literature and databases than its predecessor, and it is also implemented in Java and has an easy-to-use GUI. In addition, the tissue-specific mouse regulatory networks predicted above based on PBM and DNase data can also be directly used in DREM 2.0. Therefore, we expect DREM 2.0 to be a widely used tool in studying dynamic regulatory networks.

To conclude, this thesis first introduced new methods for finding discriminative motifs and predicting genome wide tissue-specific transcription factor binding sites. Then, the former is used to extend the DREM modeling framework (among other new features) leading to an improved tool that allows *de novo* identification of transcription factors that function in dynamic regulatory networks, and the latter provides tissue-specific binding data that can be directly used by DREM to allow more accurate modeling of tissue-specific dynamic regulatory networks when combined with expression data in the same tissue. Taken together, the thesis makes new contributions to the studying and understanding of transcriptional regulation from a computational perspective.

5.2 Future directions

Several future studies can be performed to further extend the work presented in this thesis.

5.2.1 More complicated model that allows dependence in discriminative motif finding

In DECOD, the PWM motif model is used when searching for the discriminative motif between two sets of sequences. One of the biggest disadvantages of PWM model is that it assumes independence between positions in a motif, and there has been controversies regarding whether this is oversimplied and not true in nature [6, 16, 105, 117, 204]. More complicated models than PWM to represent TF binding motifs have been proposed [163, 166, 206]. When comparing the performance of PWM-based models with our k-mer based model to infer TF binding profile from PBM data, we also showed that our model outperforms PWM-based models in predicting binding sites *in vivo* (Chapter 3). In Section we tested using a Markov motif model in DECOD on human data but it did not appear to bring much improvement over the PWM model. Still it would be interesting to study whether adopting a more complicated model such as a profile Hidden Markov model or conditional random field in place of PWM would lead to an improvement in performance for DECOD. As already discussed in Section , plugging in a more complicated model into the target function without deconvolution is straightforward, but depending on the specific model to be used, it might be difficult to incorporate the new model using the convolved motif component. Moreover, with a more complicated model it would be difficult to make the search process fast in the add step (Section 2.3.4), as the calculation of partial derivatives with respect to model parameters (Section 2.3.5) may become computationally intractable - this is one of the key steps that make DECOD faster than many other tools. In addition, with a more complicated model that allows interdependence between positions, the number of parameters to learn will increase, and therefore the motif cardinality (the number of k-mers used to construct the model, Section 2.3.7) will also need to be increased, and this will further slow down the search process. Therefore, changes in the optimization process may be necessary in order to use such more complicated models.

5.2.2 Evaluating the statistical significance of the discriminative motifs identified by DECOD

Currently DECOD only reports the target function score together with the discriminative motifs that it find. The user can only evaluate whether the reported motif is biologically meaningful by relying on its match E-value with the most similar known motif, and its relative rank among all the reported motifs. It would be very helpful if a p-value representing the significance of the discriminative motif DECOD found could be reported, with the null hypothesis being that the motif is equally enriched (or depleted) in the positive and negative sequences. However, this is not an easy task [172]. Several relevant methods have been proposed. For example, Sinha et al. [168] first introduced a method to compute p-values for discriminative motifs, but that method is prohibitively slow for large datasets. A more recent discriminative motif finding method, DREME [7], uses an enumerative approach to find regular expression motifs with Fisher's exact test p-value over a cutoff. A similar method might be considered to calculate p-values for the PWM motifs that DECOD finds.

5.2.3 More accurate methods that assign binding sites for a TF to the genes that it regulates

In many studies that predict genome wide TF binding events and regulatory networks including ours presented in Chapter 3, a specific TF is assumed to regulate a target gene if its binding site (either as a high scoring genomic region from scanning by a PWM or integrative model, or as a ChIP-seq peak) can be found within the promoter (a specific distance range from the transcription start site) of that gene [24, 27, 48, 124, 205]. Although this is simple and intuitive, it has the disadvantage that it does not take into account the distance between the binding site and the potential target gene. In addition, such an approach could not incorporate long range interactions as usually the range of the promoter region being considered is limited, yet it has been demonstrated that certain TFs can regulate genes that are over 100kbps away [112, 113]. Recently, to address the concerns mentioned above, several computational methods have been proposed to assign TF binding sites to the genes that they regulate in a more principled manner using ChIP-seq data alone [29, 130], or with the help of additional expression [14] or DNase data [112]. It would be useful to develop a similar method to assign the tissue-specific binding sites of each TF inferred from the PBM and DNase data to its target genes, as more accurate knowledge about this will directly improve the quality of the tissue-specific TRNs inferred. Experimental methods to determine long range interactions have also been developed including chromosome conformation capture carbon copy (5C) [42] and chromatin interaction analysis by paired-end tag sequencing(ChIA-PET) [53], and such information can be used to evaluate the performance of the assignment method or develop new predicative models for long range TF-gene interactions.

5.2.4 Applications to human data and the study of tissue-specific regulatory network

Our method for using protein binding microarray data has primarily focused on mouse TFs, as currently most of the available PBM data for TFs in higher eukaryotes are in mouse. However, it has been known that the binding preferences of many TFs between human and mouse are highly conserved, although individual binding events themselves may show rapid turnovers [155, 195]. Moreover, TFs that have similar DNA binding domains also usually have similar binding preferences, as we have showed using several cases in the evaluation of PLAR-PBM. Therefore, it would be interesting to transfer knowledge about the TF binding profiles learned from PBM data in mouse to corresponding TFs in human. Moreover, recently, Jolma et al. [75] used high-throughput SELEX techniques to study the binding specificities of over 800 human TFs. It would be a rich resource if such knowledge can be utilized and combined with the DNase-seq data available for hundreds of human tissue/cell types to predict a tissue-specific TF binding map for human, as such resource would be directly relevant in understanding transcriptional regulation in human and uncovering novel mechanisms in human diseases. Furthermore, although our current works enabled the study of both tissue-specific binding and dynamic regulatory networks, so far they have been two separated pieces. It would be interesting to run DREM 2.0 on a dataset where the time-series expression data in a specific tissue and the tissue-specific binding data for the same tissue is available, and compare the resulting tissue-specific regulatory network model with a non-tissue-specific model in which general binding data is used, to observe how much improvement is obtained.

Appendix A

Calculating the derivative for the deconvolved mixture component in DECOD

1. The definition of $F(\theta)$ in the convolved mixture model is

$$F(\theta) = \sum_{a \in \Sigma^k} c(a) \frac{pA}{pA + [1 - (2k-1)p]B^a} \tag{A.1}$$

$$= \sum_{a \in \Sigma^k} c(a) \left[1 - \frac{[1 - (2k-1)p]B^a}{pA + [1 - (2k-1)p]B^a} \right] \tag{A.2}$$

in which Σ^k denotes all possible k-mers, $c(a) := X(a) - Y(a)$ is the frequency difference of the k-mer a in the positive and negative sequences, p is the probability of the motif occurrance, k is the motif length, B is the background model, and

$$A := \theta^a + ([\underline{B}_1 \overline{\theta}_{k-1}]^a + \cdots + [\underline{B}_{k-1} \overline{\theta}_1]^a) + ([\underline{\theta}_1 \overline{B}_{k-1}]^a + \cdots + [\underline{\theta}_{k-1} \overline{B}_1]^a) \tag{A.3}$$

$$:= X + Y + Z \tag{A.4}$$

in which θ is a $4 \times k$ PWM matrix of the motif with columns sum to 1. We represent the k-mer a also as a $4 \times k$ matrix, each element $a_{ij} \in \{0, 1\}$ and the columns sum to 1.

$[\underline{P}_j\overline{Q}_{k-j}]$ denote the PWM obtained by taking the last j columns from the PWM P and the first $k-j$ columns from the PWM Q. We regard the background model B as a PWM also with all columns equal. We use θ^a as a shorthand for $\Pr(a|\theta)$.

2.

$$\frac{\partial F(\theta)}{\partial \theta_{mn}} = \sum_{a\in\Sigma^k} c(a)\frac{[1-(2k-1)p]B^a\cdot p}{(pA+[1-(2k-1)p]B^a)^2}\cdot\frac{\partial A}{\partial\theta_{mn}} \tag{A.5}$$

$$= p\cdot[1-(2k-1)p]\cdot\sum_{a\in\Sigma^k} c(a)\frac{B^a}{(pA+[1-(2k-1)p]B^a)^2}\cdot\frac{\partial A}{\partial\theta_{mn}} \tag{A.6}$$

$$\frac{\partial A}{\partial\theta_{mn}} = \frac{\partial X}{\partial\theta_{mn}} + \frac{\partial Y}{\partial\theta_{mn}} + \frac{\partial Z}{\partial\theta_{mn}} \tag{A.7}$$

3. For X,

$$X = \theta^a \tag{A.8}$$

$$= \prod_{i=1}^{k}\left(\sum_{j=1}^{4}\theta_{ji}a_{ji}\right) \tag{A.9}$$

$$\frac{\partial X}{\partial\theta_{mn}} = \prod_{\substack{i=1\\ i\neq n}}^{k}\left(\sum_{j=1}^{4}\theta_{ji}a_{ji}\right)\cdot a_{mn} \tag{A.10}$$

$$\tag{A.11}$$

4. For Y,

$$Y = [\underline{B}_1\overline{\theta}_{k-1}]^a + \cdots + [\underline{B}_{k-1}\overline{\theta}_1]^a \tag{A.12}$$

$$[\underline{B}_1\overline{\theta}_{k-1}]^a = \prod_{i=1}^{1}\left(\sum_{j=1}^{4}b_j a_{ji}\right)\cdot\prod_{i=2}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-1}a_{ji}\right) \tag{A.13}$$

$$[\underline{B}_2\overline{\theta}_{k-2}]^a = \prod_{i=1}^{2}\left(\sum_{j=1}^{4}b_j a_{ji}\right)\cdot\prod_{i=3}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-2}a_{ji}\right) \tag{A.14}$$

$$\vdots \tag{A.15}$$

$$[\underline{B}_{k-1}\overline{\theta}_1]^a = \prod_{i=1}^{k-1}\left(\sum_{j=1}^{4}b_j a_{ji}\right)\cdot\prod_{i=k}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-(k-1)}a_{ji}\right) \tag{A.16}$$

There are $k-1$ rows above. The last $n-1$ rows do not contain θ_{mn} so the partial derivative of them with respect to θ_{mn} for these rows will be 0. For the first $k-n$ rows,

$$\frac{\partial[\underline{B}_1\overline{\theta}_{k-1}]^a}{\partial\theta_{mn}} = \prod_{i=1}^{1}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=2\\ i\neq n+1}}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-1}a_{ji}\right) \cdot a_{m,n+1} \tag{A.17}$$

$$\frac{\partial[\underline{B}_2\overline{\theta}_{k-2}]^a}{\partial\theta_{mn}} = \prod_{i=1}^{2}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=3\\ i\neq n+2}}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-2}a_{ji}\right) \cdot a_{m,n+2} \tag{A.18}$$

$$\vdots \tag{A.19}$$

$$\frac{\partial[\underline{B}_{k-n}\overline{\theta}_n]^a}{\partial\theta_{mn}} = \prod_{i=1}^{k-n}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=k-n+1\\ i\neq k}}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-(k-n)}a_{ji}\right) \cdot a_{m,k} \tag{A.20}$$

Thus,

$$Y = \sum_{l=1}^{k-1}\left[\prod_{i=1}^{l}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{i=l+1}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-l}a_{ji}\right)\right] \tag{A.21}$$

$$\tag{A.22}$$

and

$$\frac{\partial Y}{\partial\theta_{mn}} = \sum_{l=1}^{k-n}\left[\prod_{i=1}^{l}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=l+1\\ i\neq l+n}}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-l}a_{ji}\right) \cdot a_{m,l+n}\right] \tag{A.23}$$

5. For Z,

$$Z = [\underline{\theta}_1\overline{B}_{k-1}]^a + \cdots + [\underline{\theta}_{k-1}\overline{B}_1]^a \tag{A.24}$$

$$[\underline{\theta}_1\overline{B}_{k-1}]^a = \prod_{i=1}^{1}\left(\sum_{j=1}^{4}\theta_{j,k+i-1}a_{ji}\right) \cdot \prod_{i=2}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \tag{A.25}$$

$$[\underline{\theta}_2\overline{B}_{k-2}]^a = \prod_{i=1}^{2}\left(\sum_{j=1}^{4}\theta_{j,k+i-2}a_{ji}\right) \cdot \prod_{i=3}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \tag{A.26}$$

$$\vdots \tag{A.27}$$

$$[\underline{\theta}_{k-1}\overline{B}_1]^a = \prod_{i=1}^{k-1}\left(\sum_{j=1}^{4}\theta_{j,k+i-(k-1)}a_{ji}\right) \cdot \prod_{i=k}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \tag{A.28}$$

There are $k-1$ rows above. The first $k-n$ rows do not contain θ_{mn} so the partial derivative of them with respect to θ_{mn} for these rows will be 0. For the last $n-1$ rows,

$$\frac{\partial[\underline{\theta}_{k-n+1}\overline{B}_{n-1}]^a}{\partial\theta_{mn}} = \prod_{i=k-n+2}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=1\\ i\neq 1}}^{k-n+1}\left(\sum_{j=1}^{4} \theta_{j,k-(k-n+1)+i} a_{ji}\right) \cdot a_{m,1} \tag{A.29}$$

$$\frac{\partial[\underline{\theta}_{k-n+2}\overline{B}_{n-2}]^a}{\partial\theta_{mn}} = \prod_{i=k-n+3}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=1\\ i\neq 2}}^{k-n+2}\left(\sum_{j=1}^{4} \theta_{j,k-(k-n+2)+i} a_{ji}\right) \cdot a_{m,2} \tag{A.30}$$

$$\vdots \tag{A.31}$$

$$\frac{\partial[\underline{\theta}_{k-1}\overline{B}_{1}]^a}{\partial\theta_{mn}} = \prod_{i=k}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=1\\ i\neq n-1}}^{k-1}\left(\sum_{j=1}^{4} \theta_{j,k-(k-1)+i} a_{ji}\right) \cdot a_{m,n-1} \tag{A.32}$$

Thus,

$$Z = \sum_{l=1}^{k-1}\left[\prod_{i=l+1}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{i=1}^{l}\left(\sum_{j=1}^{4} \theta_{j,k-l+i} a_{ji}\right)\right] \tag{A.33}$$

and

$$\frac{\partial Z}{\partial\theta_{mn}} = \sum_{l=k-n+1}^{k-1}\left[\prod_{i=l+1}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{\substack{i=1\\ i\neq n-k+l}}^{l}\left(\sum_{j=1}^{4} \theta_{j,k-l+i} a_{ji}\right) \cdot a_{m,n-k+l}\right] \tag{A.34}$$

6. In summary,

$$A = X + Y + Z \tag{A.35}$$

$$\begin{aligned} &= \prod_{i=1}^{k}\left(\sum_{j=1}^{4} \theta_{ji} a_{ji}\right) + \\ &\sum_{l=1}^{k-1}\left[\prod_{i=1}^{l}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{i=l+1}^{k}\left(\sum_{j=1}^{r} \theta_{j,i-l} a_{ji}\right)\right] + \\ &\sum_{l=1}^{k-1}\left[\prod_{i=l+1}^{k}\left(\sum_{j=1}^{4} b_j a_{ji}\right) \cdot \prod_{i=1}^{l}\left(\sum_{j=1}^{4} \theta_{j,k-l+i} a_{ji}\right)\right] \end{aligned} \tag{A.36}$$

and

$$\begin{aligned}\frac{\partial A}{\partial \theta_{mn}} = &\prod_{\substack{i=1\\ i\neq n}}^{k}\left(\sum_{j=1}^{4}\theta_{ji}a_{ji}\right)\cdot a_{mn}+\\ &\sum_{l=1}^{k-n}\left[\prod_{i=1}^{l}\left(\sum_{j=1}^{4}b_j a_{ji}\right)\cdot\prod_{\substack{i=l+1\\ i\neq l+n}}^{k}\left(\sum_{j=1}^{4}\theta_{j,i-l}a_{ji}\right)\cdot a_{m,l+n}\right]+\\ &\sum_{l=k-n+1}^{k-1}\left[\prod_{i=l+1}^{k}\left(\sum_{j=1}^{4}b_j a_{ji}\right)\cdot\prod_{\substack{i=1\\ i\neq n-k+l}}^{l}\left(\sum_{j=1}^{4}\theta_{j,k-l+i}a_{ji}\right)\cdot a_{m,n-k+l}\right]\end{aligned}\tag{A.37}$$

Appendix B

Full DECOD evaluation results on the 65 yeast TFs

Table B.1: Evaluation of 6 discriminative motif finding methods on recovering the 65 yeast TFs. See Section 2.7.1 for details.

TF	DC	DC-S	DC-A	DME	DEME	CMF	Sd	ALSE	N_BOUND	%Contain	W	Enrich
ABF1	+	+	+	+	+	+	+		178	86.50%	13	99
CBF1	+		+	+	+	+		+	195	68.70%	7	99
FHL1	+	+	+	+	+	+	+		131	73.30%	10	99
RAP1	+	+	+	+	+	+			109	74.30%	10	79.92
REB1	+	+	+	+	+	+	+		99	86.90%	7	77.93
UME6	+	+	+	+	+	+	+		93	68.80%	8	72.32
RPN4									70	80.00%	9	72.02
GCN4	+	+	+	+	+	+	+		143	65.00%	7	64.62
YAP7									101	80.20%	8	62.65
MCM1				+	+				77	66.20%	11	55.28

(Continued on next page...)

TF	DC	DC-S	DC-A	DME	DEME	CMF	Sd	ALSE	N_BOUND	%Contain	W	Enrich
NRG1						+			108	59.30%	7	45.42
MBP1	+	+	+	+	+	+	+		92	47.80%	7	40
SKN7								+	148	37.20%	9	38.79
CIN5	+	+	+		+	+			118	42.40%	8	38.36
SUM1	+	+	+	+	+		+		51	88.20%	10	36.47
SWI6	+	+	+	+	+	+	+		121	51.20%	7	33.62
HSF1	+	+							74	67.60%	13	32.96
SWI4	+	+	+	+	+	+	+		130	49.20%	7	31.96
TYE7	+	+	+	+	+	+	+	+	56	53.60%	8	30.56
SFP1									37	73.00%	9	26.64
FKH2	+	+			+	+	+		91	58.20%	7	26.62
HAP1					+		+		116	28.80%	11	24.72
INO4	+	+	+	+	+	+		+	32	68.80%	8	24.15
FKH1	+	+	+	+	+	+	+		104	76.90%	8	23.43
CAD1			+	+	+				29	65.50%	10	21.69
SNT2	+	+	+	+	+				20	70.00%	9	21.64
SUT1					+		+		67	37.30%	10	21.01
STE12	+	+	+		+		+		142	88.00%	7	20.86
NDD1									94	28.70%	11	20.74
LEU3				+	+				32	40.60%	10	20.45
HAP4	+	+	+	+	+	+			54	50.00%	7	20.32
AFT2									76	63.20%	6	19.4
MSN2									74	40.50%	9	18.81
PHD1						+			103	57.30%	8	17.93
YDR026c	+	+	+	+	+	+		+	15	86.70%	9	17.26
YAP1				+	+				37	51.40%	9	15.55

(Continued on next page...)

TF	DC	DC-S	DC-A	DME	DEME	CMF	Sd	ALSE	N_BOUND	%Contain	W	Enrich
THI2									49	28.60%	12	15.38
INO2	+	+	+	+	+	+			35	71.40%	7	14.97
SPT2	+	+	+						36	63.90%	11	14.4
SIP4			+						24	37.50%	13	14.36
SIG1			+		+				16	87.50%	12	13.48
STB5			+	+	+	+			44	40.90%	9	13.44
GAL4									37	32.40%	18	13.42
RDS1			+					+	49	24.50%	6	12.65
ZAP1									18	27.80%	14	12.35
SOK2	+	+	+						73	64.40%	6	12.28
MET4				+					37	21.60%	15	12.26
STB1	+	+	+	+	+	+		+	23	47.80%	9	11.95
GLN3			+						79	55.70%	7	11.65
RFX1					+		+		25	28.00%	13	11.48
AZF1			+		+				24	54.20%	18	10.85
RCS1				+					41	46.30%	7	10.44
PHO2									14	50.00%	11	10.2
IME1									36	61.10%	11	9.92
RLR1	+	+	+	+					25	64.00%	12	9.76
PDR1				+					68	22.10%	11	9.21
PHO4									24	45.80%	7	9.17
DIG1			+						66	54.50%	7	8.74
TEC1	+		+	+	+	+		+	37	78.40%	7	6.4
BAS1	+	+	+	+	+	+		+	17	52.90%	6	4.99
SPT23									45	91.10%	8	4.79
ACE2				+			+		71	28.20%	7	4.78

(Continued on next page. . .)

TF	DC	DC-S	DC-A	DME	DEME	CMF	Sd	ALSE	N_BOUND	%Contain	W	Enrich
STB4									28	21.40%	9	3.69
DAL82									62	41.90%	6	3.33
GAT1									49	36.70%	6	2.25
Sum(Top)	15	14	13	13	15	14	11	3				
Sum(All)	28	26	34	31	34	24	17	9				

DC: DECOD

DC-S: DECOD Simple without deconvolution

DC-A: DECOD with alternative seed

Sd: Seeder

+: Correclty recovered

N_BOUND: Number of probes bound by the TF in the ChIP-chip experiment

%Contain: The percentage of the bound probes containing the motif of the TF

W: The width of the motif

Enrich: The enrichment score of the motif.

Appendix C

284 mouse TFs with PBM data available

Table C.1: List of the 284 mouse TFs with PBM data available. See Section 3.2.5 for details

Alx4	Elf2	Gm397	Hoxb9	Lhx1	Nr2f2	Prrx1	Spic	Zic3
Arid3a	Elf3	Gm4881	Hoxc10	Lhx2	Obox1	Prrx2	Srf	Zif268
Arid5a	Elf4	Gm5454	Hoxc11	Lhx3	Obox2	Rara	Sry	Zscan4
Arx	Elf5	Gmeb1	Hoxc12	Lhx4	Obox3	Rax	Tbp	
Ascl2	Elk1	Gsc	Hoxc13	Lhx5	Obox5	Rfx3	Tcf1	
Atf1	Elk3	Gsh2	Hoxc4	Lhx6	Obox6	Rfx4	Tcf2	
Bapx1	Emx2	Hbp1	Hoxc5	Lhx8	Og2x	Rfxdc2	Tcf3	
Barhl1	En1	Hdx	Hoxc6	Lhx9	Osr1	Rhox11	Tcf7	
Barhl2	En2	Hic1	Hoxc8	Lmx1a	Osr2	Rhox6	Tcf7l2	
Barx1	Eomes	Hlx1	Hoxc9	Lmx1b	Otp	Rxra	Tcfap2a	
Barx2	Erg	Hlxb9	Hoxd1	Mafb	Otx1	Sfpi1	Tcfap2b	
Bbx	Esrra	Hmbox1	Hoxd10	Mafk	Otx2	Shox2	Tcfap2c	
Bcl6b	Esx1	Hmx1	Hoxd11	Max	Pax4	Six1	Tcfap2e	

(Continued on next page. . .)

Bhlhb2
Bsx
Cart1
Cdx2
Cphx
Crx
Cutl1
Dbx1
Dbx2
Dlx1
Dlx2
Dlx3
Dlx4
Dlx5
Dmbx1
Dobox4
Dobox5
Duxl
E2F2
E2F3
Egr1
Ets1
Etv1
Etv3
Etv4
Etv5
Etv6
Evx1
Evx2
Fli1
Foxa2
Foxj1
Foxj3
Foxk1
Foxl1
Gabpa
Gata3
Gata5
Gata6
Gbx1
Gbx2
Gcm1
Hmx2
Hmx3
Hnf4a
Homez
Hoxa1
Hoxa10
Hoxa11
Hoxa13
Hoxa2
Hoxa3
Hoxa4
Hoxa5
Hoxa6
Hoxa7
Hoxa9
Hoxb13
Hoxb3
Hoxb4
Hoxb5
Hoxb6
Hoxb7
Hoxd12
Hoxd13
Hoxd3
Hoxd8
IRC900814
Ipf1
Irf3
Irf4
Irf5
Irf6
Irx2
Irx3
Irx4
Irx5
Irx6
Isgf3g
Isl2
Isx
Jundm2
Klf7
Lbx2
Meis1
Meox1
Mrg1
Mrg2
Msx1
Msx2
Msx3
Mtf1
Myb
Mybl1
Myf6
Nkx1-1
Nkx1-2
Nkx2-2
Nkx2-3
Nkx2-4
Nkx2-5
Nkx2-6
Nkx2-9
Nkx3-1
Nkx6-1
Pax6
Pax7
Pbx1
Phox2a
Phox2b
Pitx1
Pitx2
Pitx3
Pknox1
Pknox2
Plagl1
Pou1f1
Pou2f1
Pou2f2
Pou2f3
Pou3f1
Pou3f2
Pou3f3
Pou3f4
Pou4f3
Pou6f1
Six2
Six3
Six4
Six6
Smad3
Sox1
Sox11
Sox12
Sox13
Sox14
Sox15
Sox17
Sox18
Sox21
Sox30
Sox4
Sox5
Sox7
Sox8
Sp100
Sp4
Tcfe2a
Tgif1
Tgif2
Titf1
Tlx2
Uncx4
Vax1
Vax2
Vsx1
Zbtb12
Zbtb3
Zbtb7b
Zfp105
Zfp128
Zfp161
Zfp187
Zfp281
Zfp410
Zfp691
Zfp740
Zic1

Appendix D

55 mouse tissue/cell types with DNase data available

Table D.1: List of the 55 mouse tissue/cell types with DNase data available. See Section 3.3.1 for details. Descriptions for the tissue/cell types are from the ENCODE project website

Name	Description	Category	Tissue
3134RiiiMImmortal	Mammary	cellLine	mammary
416bC57bl6MAdult8wks	myeloid progenitor cells, CD34+	cellLine	blood
A20BalbcannMAdult8wks	B cell lymphoma line derived from a spontaneous reticulum cell neoplasm	cellLine	blood
Bcellcd19pC57bl6MAdult8wks	B Cell , CD19+	primaryCells	blood
Bcellcd43nC57bl6MAdult8wks	mouse spleen B cells, CD43-,CD11b-	primaryCells	blood
CerebellumC57bl6MAdult8wks	Cerebellum	Tissue	cerebellum
CerebrumC57bl6MAdult8wks	Cerebrum	Tissue	cerebrum
Ch122a4bFImmortal	B-cell lymphoma (GM12878 analog)	cellLine	blood
EpcmppCd1ME14half	liver fraction CD117-,CD71+,TER119+	primaryCell	blood

(Continued on next page. . .)

Name	Description	Category	Tissue
EpcpmmCd1ME14half	liver fraction CD117+,CD71-,TER119-	primaryCell	blood
EpcppmCd1ME14half	liver fraction CD117+,CD71+,TER119-	primaryCell	blood
EpcpppCd1ME14half	liver fraction CD117+,CD71+,TER119+	primaryCell	blood
Escj7S129ME0	ES-cells were originally isolated from 129S1/SVImJ mice [180]	primaryCells	
Ese14129olaME0	mouse embryonic stem cell line E14	primaryCells	
Esww6koUknME0	Histone H1c, H1d, H1e triple null mouse embryonic stem cell line derived from ES-WW6 cells.	primaryCells	
Esww6UknME0	ES-cells isolated from mix of 20% C57/B6J, 75% 129/Sv and 5% SJL strains	primaryCells	
FatC57bl6MAdult8wks	Adipose tissue	Tissue	adipose
FibroblastC57bl6MAdult8wks	Fibroblast	Tissue	lung
FlbudCd1ME11half	embryo forelimb buds	Tissue	limb
GfatC57bl6MAdult8wks	Genital Adipose tissue	Tissue	adipose
HeartC57bl6MAdult8wks	Heart	Tissue	heart
HlbudCd1ME11half	embryo hindlimb buds	Tissue	limb
HlembryoCd1ME11half	Whole embryos with heads removed	Tissue	embryo
KidneyC57bl6MAdult8wks	Kidney	Tissue	kidney
LgintC57bl6MAdult8wks	Large Intestine	Tissue	large intestine
Liver129dlcrME14half	Liver	Tissue	liver
LiverC57bl6MAdult8wks	Liver	Tissue	liver
LiverC57bl6ME14half	Liver	Tissue	liver
LiverS129ME14half	Liver	Tissue	liver
LungC57bl6MAdult8wks	Lung	Tissue	lung
MelC57bl6MAdult8wks	Leukemia (K562 analog)	cellLine	blood

(Continued on next page. . .)

Name	Description	Category	Tissue
MesodermCd1ME11half	axial somatic and lateral plate mesoderm from eviscerated headless, limbless embryos	Tissue	mesoderm
MgerUknImmortalDiffc24h	MEL-GATA-1-ER, This is a mouse suspension cell line derived from MEL cells by stable transfection with a GATA-1-ER fusion protein construct [30]. These cells can be terminally differentiated into mature erythroid cells with β-estradiol treatment	cellLine	blood
MgerUknImmortalDiffc48h	MEL-GATA-1-ER, Same as above	cellLine	blood
MgerUknImmortal	MEL-GATA-1-ER, Same as above	cellLine	blood
Nih3t3NihsMImmortal	fibroblast	cellLine	blood
PatskiSpbl6MImmortal	Mouse Embryonic Kidney Fibroblast	cellLine	kidney
RetinaC57bl6MAdult1wks	Retina	Tissue	brain
RetinaC57bl6MAdult8wks	Retina	Tissue	brain
RetinaC57bl6MNew1days	Retina	Tissue	brain
SkmuscleC57bl6MAdult8wks	Skeletal Muscle	Tissue	skeletal muscle
SpleenC57bl6MAdult8wks	Spleen	Tissue	spleen
ThelpaC57bl6MAdult8wks	Activated primary CD4 effector cells, isolated ex vivo	primaryCells	blood
ThymusC57bl6MAdult8wks	Thymus	Tissue	thymus
TnaiveC57bl6MAdult8wks	Thymus	Tissue	thymus
TregaC57bl6MAdult8wks	Activated primary T regulatory cells, isolated ex vivo	primaryCells	blood
TregC57bl6MAdult8wks	Regulatory T cells CD4+,CD25+	primaryCells	blood
WbrainC57bl6MAdult8wks	Whole Brain	Tissue	brain
WbrainC57bl6ME14half	Whole Brain	Tissue	brain

(Continued on next page…)

Name	Description	Category	Tissue
WbrainC57bl6ME18half	Whole Brain	Tissue	brain
WholebrainC57bl6MAdult8wks	Whole Brain	Tissue	brain
WholebrainC57bl6ME14half	Whole Brain	Tissue	brain
Zhbtc4129olaME0Diffb24h	ndifferentiated mouse embryonic stem cells	primaryCells	
Zhbtc4129olaME0Diffb6h	ndifferentiated mouse embryonic stem cells	primaryCells	
Zhbtc4129olaME0	ndifferentiated mouse embryonic stem cells	primaryCells	

Bibliography

[1] Phaedra Agius, Aaron Arvey, William Chang, William Stafford Noble, and Christina Leslie. High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Comp Biol*, 6(9), 2010. (Cited in 1.2.2, 5, 3.6.2, 3.6.2, 3.7)

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Sciences, 5th edition, November 2007. (Cited in 1.1)

[3] Jose M Alonso, Anna N Stepanova, Thomas J Leisse, Christopher J Kim, Huaming Chen, Paul Shinn, Denise K Stevenson, Justin Zimmerman, Pascual Barajas, Rosa Cheuk, Carmelita Gadrinab, Collen Heller, Albert Jeske, Eric Koesema, Cristina C Meyers, Holly Parker, Lance Prednis, Yasser Ansari, Nathan Choy, Hashim Deen, Michael Geralt, Nisha Hazari, Emily Hom, Meagan Karnes, Celene Mulholland, Ral Ndubaku, Ian Schmidt, Plinio Guzman, Laura Aguilar-Henonin, Markus Schmid, Detlef Weigel, David E Carter, Trudy Marchand, Eddy Risseeuw, Debra Brogden, Albana Zeko, William L Crosby, Charles C Berry, and Joseph R Ecker. Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, 301(5633):653–657, August 2003. (Cited in 4.3.10)

[4] Jose M Alonso, Anna N Stepanova, Roberto Solano, Ellen Wisman, Simone Ferrari, Frederick M Ausubel, and Joseph R Ecker. Five components of the ethylene-response pathway identified in a screen for weak ethylene-insensitive mutants in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.*, 100(5): 2992–2997, March 2003. (Cited in 4.3.1)

[5] Matti Annala, Kirsti Laurila, Harri Lähdesmäki, and Matti Nykter. A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS ONE*, 6(5):e20059, 2011. (Cited in 1.2.2)

[6] Gwenael Badis, Michael F Berger, Anthony A Philippakis, Shaheynoor Talukder, Andrew R Gehrke, Savina A Jaeger, Esther T Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, Hanna Kuznetsov, Chi-Fong Wang, David Coburn, Daniel E Newburger, Quaid Morris, Timothy R Hughes, and Martha L Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, June 2009. (Cited in 1.2.2, 1.2.5, 2.9, 3.1, 3.6.1, 3.6.2, 3.7, 5.2.1)

[7] Timothy L Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, June 2011. (Cited in 5.2.2)

[8] Timothy L Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994. (Cited in 2.1)

[9] Mukesh Bansal, Giusy Della Gatta, and Diego di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7): 815–822, April 2006. (Cited in 4.1.1)

[10] Ziv Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–2503, November 2004. (Cited in 4.1.1)

[11] Ziv Bar-Joseph, Georg Gerber, Itamar Simon, David K Gifford, and Tommi S Jaakkola. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc. Natl. Acad. Sci. U.S.A.*, 100(18):10146–10151, September 2003. (Cited in 4.1.1)

[12] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.*, 13(8):552–564, August 2012. (Cited in 4.1.1)

[13] Andreas S Barth, Ruprecht Kuner, Andreas Buness, Markus Ruschhaupt, Sylvia Merk, Ludwig Zwermann, Stefan Kääb, Eckart Kreuzer, Gerhard Steinbeck, Ulrich Mansmann, Annemarie Poustka, Michael Nabauer, and Holger Sültmann. Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *J. Am. Coll. Cardiol.*, 48(8): 1610–1617, October 2006. (Cited in 1.1)

[14] Tobias Bauer, Roland Eils, and Rainer König. RIP: the regulatory interaction predictor–a machine

learning-based approach for predicting target genes of transcription factors. *Bioinformatics*, 27 (16):2239–2247, August 2011. (Cited in 5.2.3)

[15] Yoshua Bengio and Paolo Frasconi. An input output HMM architecture. *Advances in neural information processing systems*, pages 427–434, 1995. (Cited in 4.1.1)

[16] Panayiotis V Benos, Martha L Bulyk, and Gary D Stormo. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, 30(20):4442–4451, October 2002. (Cited in 5.2.1)

[17] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, 24(11):1429–1435, November 2006. (Cited in 1.2.2, 1.1, 1.2.2, 3.1, 3.2.2, 3, 3.6.1, 3.6.2, 3.7)

[18] Brad M Binder, Laura A Mortimore, Anna N Stepanova, Joseph R Ecker, and Anthony B Bleecker. Short-term growth responses to ethylene in Arabidopsis seedlings are EIN3/EIL1 independent. *Plant Physiol.*, 136(2):2921–2927, October 2004. (Cited in 4.1.2, 4.4.2)

[19] Valentina Boeva, Didier Surdez, Noëlle Guillon, Franck Tirode, Anthony P Fejes, Olivier Delattre, and Emmanuel Barillot. De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.*, 38(11):e126, June 2010. (Cited in 2.1)

[20] Sylvia F Boj, Johan H van Es, Meritxell Huch, Vivian S W Li, Anabel José, Pantelis Hatzis, Michal Mokry, Andrea Haegebarth, Maaike van den Born, Pierre Chambon, Peter Voshol, Yuval Dor, Edwin Cuppen, Cristina Fillat, and Hans Clevers. Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand. *Cell*, 151(7):1595–1607, December 2012. (Cited in 3.6.5)

[21] E Boy-Marcotte, G Lagniel, M Perrot, F Bussereau, A Boudsocq, M Jacquet, and J Labarre. The heat shock response in yeast: differential regulations and contributions of the Msn2p/Msn4p and Hsf1p regulons. *Mol. Microbiol.*, 33(2):274–283, July 1999. (Cited in 1.1)

[22] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle

da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, 36(Database issue):D102–6, January 2008. (Cited in 2.9)

[23] Michael J Buck and Jason D Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, March 2004. (Cited in 1.2.4, 3.1)

[24] Katherine N Chang, Shan Zhong, Matthew T Weirauch, Gary C Hon, Mattia Pelizzola, Hai Li, Shao-shan Carol Huang, Robert J Schmitz, Mark A Urich, Dwight Kuo, Joseph Nery, Hong Qiao, Ally Yang, Abdullah Jamali, Trey Ideker, Bing Ren, Ziv Bar-Joseph, Timothy R Hughes, and Joseph R Ecker. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. *Submitted*. (Cited in 1.2.2, 1.3, 5.2.3)

[25] Qimin Chao, Madge Rothenberg, Roberto Solano, Gregg Roman, William Terzaghi, and Joseph R Ecker. Activation of the ethylene gas response pathway in Arabidopsis by the nuclear protein ETHYLENE-INSENSITIVE3 and related proteins. *Cell*, 89(7):1133–1144, June 1997. (Cited in 4.3.1, 4.4.1)

[26] Danian Chen, Marek Pacal, Pamela Wenzel, Paul S Knoepfler, Gustavo Leone, and Rod Bremner. Division and apoptosis of E2f-deficient retinal progenitors. *Nature*, 462(7275):925–929, December 2009. (Cited in 3.3)

[27] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B Vega, Eleanor Wong, Yuriy L Orlov, Weiwei Zhang, Jianming Jiang, Yuin-Han Loh, Hock Chuan Yeo, Zhen Xuan Yeo, Vipin Narang, Kunde Ramamoorthy Govindarajan, Bernard Leong, Atif Shahab, Yijun Ruan, Guillaume Bourque, Wing-Kin Sung, Neil D Clarke, Chia-Lin Wei, and Huck-Hui Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6): 1106–1117, June 2008. (Cited in 2.1, 3.1, 3.1, 5.2.3)

[28] Xinbin Chen, Linda J Ko, Lata Jayaraman, and Carol Prives. p53 levels, functional domains, and DNA damage determine the extent of the apoptotic response of tumor cells. *Genes Dev.*, 10(19): 2438–2451, October 1996. (Cited in 2.5.2)

[29] Chao Cheng, Renqiang Min, and Mark Gerstein. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics*, 27(23):3221–3227, December 2011. (Cited in 5.2.3)

[30] Kevin S Choe, Farshid Radparvar, Igor Matushansky, Natasha Rekhtman, Xing Han, and Arthur I Skoultchi. Reversal of tumorigenicity and the block to differentiation in erythroleukemia cells by GATA-1. *Cancer Res.*, 63(19):6363–6369, October 2003. (Cited in D.1)

[31] Ana Conesa, María José Nueda, Alberto Ferrer, and Manuel Talón. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9):1096–1102, May 2006. (Cited in 4.1.1)

[32] Joseph C Corbo, Karen A Lawrence, Marcus Karlstetter, Connie A Myers, Musa Abdelaziz, William Dirkes, Karin Weigelt, Martin Seifert, Vladimir Benes, Lars G Fritsche, Bernhard H F Weber, and Thomas Langmann. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Research*, 20(11):1512–1525, November 2010. (Cited in 3.1, 3.6.4, 3.3)

[33] Lynn M Corcoran and Maria Karvelas. Oct-2 is required early in T cell-independent B cell activation for G1 progression and for proliferation. *Immunity*, 1(8):635–645, November 1994. (Cited in 3.3)

[34] Ivan G Costa, Stefan Roepcke, Christoph Hafemeister, and Alexander Schliep. Inferring differentiation pathways from gene expression. *Bioinformatics*, 24(13):i156–64, July 2008. (Cited in 4.1.1)

[35] Gregory E Crawford, Ingeborg E Holt, James Whittle, Bryn D Webb, Denise Tai, Sean Davis, Elliott H Margulies, Yidong Chen, John A Bernat, David Ginsburg, Daixing Zhou, Shujun Luo, Thomas J Vasicek, Mark J Daly, Tyra G Wolfsberg, and Francis S Collins. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, 16(1):123–131, January 2006. (Cited in 1.2.6)

[36] Gabriel Cuellar-Partida, Fabian A Buske, Robert C McLeay, Tom Whitington, William Stafford Noble, and Timothy L Bailey. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1):56–62, January 2012. (Cited in 3.1)

[37] Jun Dai, Cai Zhang, Zhigang Tian, and Jian Zhang. Expression profile of HMBOX1, a novel transcription factor, in human cancers using highly specific monoclonal antibodies. *Exp Ther Med*, 2(3):487–490, May 2011. (Cited in 3.6.5)

[38] Modan K Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8 Suppl 7:S21, 2007. (Cited in 2.1)

[39] E R DeLong, D M DeLong, and D L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 1988. (Cited in 3.6.2, 3.2)

[40] Patrik D'haeseleer. What are DNA sequence motifs? *Nat. Biotechnol.*, 24(4):423–425, April 2006. (Cited in 1.2.5, 1.3, 2.1)

[41] David Dornan, Mirjam Eckert, Maura Wallace, Harumi Shimizu, Eleanor Ramsay, Ted R Hupp, and Kathryn L Ball. Interferon regulatory factor 1 binding to p300 stimulates DNA-dependent acetylation of p53. *Mol. Cell. Biol.*, 24(22):10083–10098, November 2004. (Cited in 2.8)

[42] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D Green, and Job Dekker. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, October 2006. (Cited in 5.2.3)

[43] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, April 2004. (Cited in 3.2.3)

[44] ENCODE Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, Michael S Kuehn, Christopher M Taylor, Shane Neph, Christoph M Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A Greenbaum, Robert M Andrews, Paul Flicek, Patrick J Boyle, Hua Cao, Nigel P Carter, Gayle K Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C Dillon, Michael O Dorschner, Heike Fiegler, Paul G Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D

James, Brett E Johnson, Ericka M Johnson, Tristan T Frum, Elizabeth R Rosenzweig, Neerja Karnani, Kirsten Lee, Gregory C Lefebvre, Patrick A Navas, Fidencio Neri, Stephen C J Parker, Peter J Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S Collins, Job Dekker, Jason D Lieb, Thomas D Tullius, Gregory E Crawford, Shamil Sunyaev, William S Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill Cheng, Heather A Hirsch, Edward A Sekinger, Julien Lagarde, Josep F Abril, Atif Shahab, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C Dickson, Daryl J Thomas, Matthew T Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, XiaoDong Zhao, K G Srinivasan, Wing-Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter, Michael L Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G Clark, James B Brown, Madhavan Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M Myers, Jane Rogers, Peter F Stadler, Todd M Lowe, Chia-Lin Wei, Yijun Ruan, Kevin Struhl, Mark Gerstein, Stylianos E Antonarakis, Yutao Fu, Eric D Green, Ulaş Karaöz, Adam Siepel, James Taylor, Laura A Liefer, Kris A Wetterstrand, Peter J Good, Elise A Feingold, Mark S Guyer, Gregory M Cooper, George Asimenos, Colin N Dewey, Minmei Hou, Sergey Nikolaev, Juan I Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R Zhang, Ian Holmes, James C Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W James Kent, Eric A Stone, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Serafim Batzoglou, Nick Goldman, Ross C Hardison, David Haussler, Webb Miller, Arend

Sidow, Nathan D Trinklein, Zhengdong D Zhang, Leah Barrera, Rhona Stuart, David C King, Adam Ameur, Stefan Enroth, Mark C Bieda, Jonghwan Kim, Akshay A Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B Vega, Charlie W H Lee, Patrick Ng, Atif Shahab, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J Oberley, David Inman, Michael A Singer, Todd A Richmond, Kyle J Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C Fowler, Phillippe Couttet, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Cordelia F Langford, David A Nix, Ghia Euskirchen, Stephen Hartman, and Al... Urban. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447 (7146):799–816, June 2007. (Cited in 2.9)

[45] Jason Ernst and Ziv Bar-Joseph. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7:191, 2006. (Cited in 4.1.1)

[46] Jason Ernst, Oded Vainas, Christopher T Harbison, Itamar Simon, and Ziv Bar-Joseph. Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.*, 3:74, 2007. (Cited in 1.3, 2.1, 3.7, 4, 4.1.1, 4.5)

[47] Jason Ernst, Qasim K Beg, Krin A Kay, Gábor Balázsi, Zoltán N Oltvai, and Ziv Bar-Joseph. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS Comp Biol*, 4(3):e1000044, March 2008. (Cited in 4.1.1)

[48] Jason Ernst, Heather L Plasterer, Itamar Simon, and Ziv Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research*, 20(4): 526–536, April 2010. (Cited in 3.1, 3.7, 5.2.3)

[49] François Fauteux, Mathieu Blanchette, and Martina V Strömvik. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, 24(20):2303–2307, October 2008. (Cited in 2.1, 2.2.6, 2.3, 2.4.1, 2.6)

[50] Barrett C Foat, Alexandre V Morozov, and Harmen J Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14): e141–9, July 2006. (Cited in 3.2.6)

[51] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303

(5659):799–805, February 2004. (Cited in 4.1.1)

[52] Martin C Frith, Ulla Hansen, John L Spouge, and Zhiping Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.*, 32(1):189–200, 2004. (Cited in 2.1)

[53] Melissa J Fullwood, Chia-Lin Wei, Edison T Liu, and Yijun Ruan. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research*, 19(4): 521–532, April 2009. (Cited in 5.2.3)

[54] Philip J Gage, Hoonkyo Suh, and Sally A Camper. Dosage requirement of Pitx2 for development of multiple organs. *Development*, 126(20):4643–4651, October 1999. (Cited in 3.3)

[55] David J Galas and Albert Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, 5(9):3157–3170, September 1978. (Cited in 3.4.2)

[56] Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241–4257, December 2000. (Cited in 4.1.1)

[57] Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011. (Cited in 3.4.2)

[58] Struan F A Grant, Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Andrei Manolescu, Jesus Sainz, Agnar Helgason, Hreinn Stefansson, Valur Emilsson, Anna Helgadottir, Unnur Styrkarsdottir, Kristinn P Magnusson, G Bragi Walters, Ebba Palsdottir, Thorbjorg Jonsdottir, Thorunn Gudmundsdottir, Arnaldur Gylfason, Jona Saemundsdottir, Robert L Wilensky, Muredach P Reilly, Daniel J Rader, Yu Bagger, Claus Christiansen, Vilmundur Gudnason, Gunnar Sigurdsson, Unnur Thorsteinsdottir, Jeffrey R Gulcher, Augustine Kong, and Kari Stefansson. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.*, 38(3):320–323, March 2006. (Cited in 3.6.5)

[59] David S Gross and William T Garrard. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, 57:159–197, 1988. (Cited in 1.2.6, 3.6.3)

[60] Christian A Grove, Federico De Masi, M Inmaculada Barrasa, Daniel E Newburger, Mark J Alkema, Martha L Bulyk, and Albertha J M Walhout. A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell*, 138(2):314–327, July 2009. (Cited in 1.2.2, 3.1)

[61] Fei Gu, Hang-Kai Hsu, Pei-Yin Hsu, Jiejun Wu, Yilin Ma, Jeffrey Parvin, Tim H-M Huang, and Victor X Jin. Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. *BMC Syst Biol*, 4:170, 2010. (Cited in 4.1.1)

[62] Hongwei Guo and Joseph R Ecker. Plant responses to ethylene gas are mediated by SCF(EBF1/EBF2)-dependent proteolysis of EIN3 transcription factor. *Cell*, 115(6):667–677, December 2003. (Cited in 4.1.2, 4.3.3, 4.4.1)

[63] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D MacIsaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, September 2004. (Cited in 2.1, 2.4.2, 2.4.3, 2.7.1, 3.1)

[64] Graham P Hayhurst, Ying-Hue Lee, Gilles Lambert, Jerrold M Ward, and Frank J Gonzalez. Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. *Mol. Cell. Biol.*, 21(4):1393–1403, February 2001. (Cited in 3.6.4, 3.3)

[65] Aibin He, Sek Won Kong, Qing Ma, and William T Pu. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl. Acad. Sci. U.S.A.*, 108(14):5632–5637, April 2011. (Cited in 3.1)

[66] Xin He, Chieh-Chun Chen, Feng Hong, Fang Fang, Saurabh Sinha, Huck-Hui Ng, and Sheng Zhong. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS ONE*, 4(12):e8155, 2009. (Cited in 3.2.6)

[67] Anne K Hennig, Guang-Hua Peng, and Shiming Chen. Regulation of photoreceptor gene expres-

sion by Crx-associated transcription factor network. *Brain Research*, 1192:114–133, February 2008. (Cited in 3.6.4)

[68] Ming Hu, Jindan Yu, Jeremy M G Taylor, Arul M Chinnaiyan, and Zhaohui S Qin. On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, 38 (7):2154–2167, April 2010. (Cited in 2.1)

[69] Yuhui Hu, Hongxia Sun, Jeffrey Drake, Frances Kittrell, Martin C Abba, Li Deng, Sally Gaddis, Aysegul Sahin, Keith Baggerly, Daniel Medina, and C Marcelo Aldaz. From mice to humans: identification of commonly deregulated genes in mammary cancer via comparative SAGE studies. *Cancer Res.*, 64(21):7748–7755, November 2004. (Cited in 1.1)

[70] Peter Huggins, Shan Zhong, Idit Shiff, Rachel Beckerman, Oleg Laptenko, Carol Prives, Marcel H Schulz, Itamar Simon, and Ziv Bar-Joseph. DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics*, 27(17):2361–2367, September 2011. (Cited in 1.3, 4.1.1, 4.2.1, 4.5)

[71] Katsumi Iizuka and Yukio Horikawa. Regulation of lipogenesis via BHLHB2/DEC1 and ChREBP feedback looping. *Biochem. Biophys. Res. Commun.*, 374(1):95–100, September 2008. (Cited in 3.6.5)

[72] Hisakazu Iwama, Tsutomu Masaki, and Shigeki Kuriyama. Abundance of microRNA target motifs in the 3'-UTRs of 20527 human genes. *FEBS Lett.*, 581(9):1805–1810, May 2007. (Cited in 2.1)

[73] Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, 43(3):264–268, March 2011. (Cited in 3.1, 3.6.3)

[74] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, June 2007. (Cited in 2.1, 3.1)

[75] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-Binding Specificities of Human Transcription Factors.

Cell, 152(1-2):327–339, January 2013. (Cited in 5.2.4)

[76] Tommy Kaplan, Xiao-Yong Li, Peter J Sabo, Sean Thomas, John A Stamatoyannopoulos, Mark D Biggin, and Michael B Eisen. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet.*, 7(2): e1001290, 2011. (Cited in 3.7)

[77] O Karni-Schmidt, A Friedler, A Zupnick, K McKinney, M Mattia, R Beckerman, P Bouvet, M Sheetz, A Fersht, and C Prives. Energy-dependent nucleolar localization of p53 in vitro requires two discrete regions within the p53 carboxyl terminus. *Oncogene*, 26(26):3878–3891, May 2007. (Cited in 2.8)

[78] Kerstin Kaufmann, Jose M Muiño, Ruy Jauregui, Chiara A Airoldi, Cezary Smaczniak, Pawel Krajewski, and Gerco C Angenent. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. *PLoS Biol.*, 7(4): e1000090, April 2009. (Cited in 3.1)

[79] Mandy D Kendrick and Caren Chang. Ethylene signaling: new levels of complexity and regulation. *Current Opinion in Plant Biology*, 11(5):479–485, October 2008. (Cited in 4.1.2)

[80] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, 26(12):1351–1359, December 2008. (Cited in 4.3.7)

[81] Konstantin Khetchoumian, Marius Teletin, Johan Tisserand, Manuel Mark, Benjamin Herquel, Mihaela Ignat, Jessica Zucman-Rossi, Florence Cammas, Thierry Lerouge, Christelle Thibault, Daniel Metzger, Pierre Chambon, and Régine Losson. Loss of Trim24 (Tif1alpha) gene function confers oncogenic activity to retinoic acid receptor alpha. *Nat. Genet.*, 39(12):1500–1506, December 2007. (Cited in 3.3)

[82] Farhad Khosrowshahian, Marian Wolanski, Wing Y Chang, Kazuhiro Fujiki, Larry Jacobs, and Michael J Crawford. Lens and retina formation require expression of Pitx3 in Xenopus pre-lens ectoderm. *Dev. Dyn.*, 234(3):577–589, November 2005. (Cited in 3.3)

[83] Kentaro Kojima, Akemi Takata, Charles Vadnais, Motoyuki Otsuka, Takeshi Yoshikawa, Masao

Akanuma, Yuji Kondo, Young Jun Kang, Takahiro Kishikawa, Naoya Kato, Zhifang Xie, Weiping J Zhang, Haruhiko Yoshida, Masao Omata, Alain Nepveu, and Kazuhiko Koike. MicroRNA122 is a key regulator of -fetoprotein expression and influences the aggressiveness of hepatocellular carcinoma. *Nat Commun*, 2:338, 2011. (Cited in 3.6.5)

[84] Jan-Philipp Kruse and Wei Gu. Modes of p53 regulation. *Cell*, 137(4):609–622, May 2009. (Cited in 2.8)

[85] H Daniel Lacorazza, Yasushi Miyazaki, Antonio Di Cristofano, Anthony Deblasio, Cyrus Hedvat, Jin Zhang, Carlos Cordon-Cardo, Shifeng Mao, Pier Paolo Pandolfi, and Stephen D Nimer. The ETS protein MEF plays a critical role in perforin gene expression and the development of natural killer and NK-T cells. *Immunity*, 17(4):437–449, October 2002. (Cited in 3.3)

[86] Eva LaTulippe, Jaya Satagopan, Alex Smith, Howard Scher, Peter Scardino, Victor Reuter, and William L Gerald. Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res.*, 62(15):4499–4506, August 2002. (Cited in 1.1)

[87] Charles E Lawrence, Stephen F Altschul, Mark S Boguski, Jun S Liu, Andrew F Neuwald, and John C Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, October 1993. (Cited in 2.1)

[88] Tong Ihn Lee, Sarah E Johnstone, and Richard A Young. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc*, 1(2):729–748, 2006. (Cited in 2.5.4)

[89] Wei-Po Lee and Wen-Shyong Tzou. Computational methods for discovering gene networks from expression data. *Brief. Bioinformatics*, 10(4):408–423, July 2009. (Cited in 4.1.1)

[90] Jeffrey T Leek, Eva Monsen, Alan R Dabney, and John D Storey. EDGE: extraction and analysis of differential gene expression. *Bioinformatics*, 22(4):507–508, February 2006. (Cited in 4.1.1)

[91] Anne Lehman, Robert Black, and Joseph R Ecker. HOOKLESS1, an ethylene response gene, is required for differential cell elongation in the Arabidopsis hypocotyl. *Cell*, 85(2):183–194, April 1996. (Cited in 4.3.1)

[92] Ordan J Lehmann, Jane C Sowden, Peter Carlsson, Tim Jordan, and Shomi S Bhattacharya. Fox's

in development and disease. *Trends Genet.*, 19(6):339–344, June 2003. (Cited in 1.1)

[93] Henry C M Leung and Francis Y L Chin. Finding motifs from all sequences with and without binding sites. *Bioinformatics*, 22(18):2217–2223, September 2006. (Cited in 2.1, 2.2.5, 2.3, 2.4.1, 2.6)

[94] Huai Li and Ming Zhan. Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data. *Bioinformatics*, 24(17):1874–1880, September 2008. (Cited in 3.7)

[95] Xiao-Yong Li, Sean Thomas, Peter J Sabo, Michael B Eisen, John A Stamatoyannopoulos, and Mark D Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biology*, 12(4):R34, 2011. (Cited in 3.1, 3.6.3)

[96] James C Liao, Riccardo Boscolo, Young-Lyeol Yang, Linh My Tran, Chiara Sabatti, and Vwani P Roychowdhury. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. U.S.A.*, 100(26):15522–15527, December 2003. (Cited in 4.1.1)

[97] Tien-ho Lin, Robert F Murphy, and Ziv Bar-Joseph. Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM Trans Comput Biol Bioinform*, 8(2):441–451, March 2011. (Cited in 2.1)

[98] Chaim Linhart, Yonit Halperin, and Ron Shamir. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Research*, 18(7): 1180–1189, July 2008. (Cited in 2.1)

[99] Zachary Lippman, Anne-Valérie Gendrel, Vincent Colot, and Rob Martienssen. Profiling DNA methylation patterns using genomic tiling microarrays. *Nat. Methods*, 2(3):219–224, March 2005. (Cited in 4.3.3)

[100] X D Liu, P C Liu, N Santoro, and D J Thiele. Conservation of a stress response: human heat shock transcription factors functionally substitute for yeast HSF. *EMBO J.*, 16(21):6466–6477, November 1997. (Cited in 1.1)

[101] Xiaole Liu, Douglas L Brutlag, and Jun S Liu. BioProspector: discovering conserved DNA motifs

in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–138, 2001. (Cited in 2.1)

[102] Belén Lucas, Karen Grigo, Silke Erdmann, Jörn Lausen, Ludger Klein-Hitpass, and Gerhart U Ryffel. HNF4alpha reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma. *Oncogene*, 24(42):6418–6431, September 2005. (Cited in 3.6.4)

[103] Kyle L MacQuarrie, Abraham P Fong, Randall H Morse, and Stephen J Tapscott. Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet.*, 27(4):141–148, April 2011. (Cited in 4.4.1)

[104] Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, August 2005. (Cited in 4.4)

[105] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, January 2007. (Cited in 1.2.5, 2.9, 3.1, 5.2.1)

[106] Shaun Mahony and Panayiotis V Benos. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, 35(Web Server issue):W253–8, July 2007. (Cited in 2.7.1, 2.8, 2.11, 2.8, 4.2.1, 5.1)

[107] Daniel Marbach, Sushmita Roy, Ferhat Ay, Patrick E Meyer, Rogerio Candeias, Tamer Kahveci, Christopher A Bristow, and Manolis Kellis. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome Research*, 22(7):1334–1349, July 2012. (Cited in 3.7)

[108] Florian Markowetz and Rainer Spang. Inferring cellular networks–a review. *BMC Bioinformatics*, 8 Suppl 6:S5, 2007. (Cited in 4.1.1)

[109] Mike J Mason, Kathrin Plath, and Qing Zhou. Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics*, 26(22):2826–2832, November 2010. (Cited in 2.1, 2.2.4, 2.4.1, 2.6)

[110] Olivier Mathieu, Zuzana Jasencakova, Isabelle Vaillant, Anne-Valérie Gendrel, Vincent Colot, Ingo

Schubert, and Sylvette Tourmente. Changes in 5S rDNA chromatin organization and transcription during heterochromatin establishment in Arabidopsis. *Plant Cell*, 15(12):2929–2939, December 2003. (Cited in 4.3.3)

[111] V Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, Nico Voss, Philip Stegmaier, Birgit Lewicki-Potapov, H Saxel, Alexander E Kel, and Edgar Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34(Database issue):D108–10, January 2006. (Cited in 2.4.3, 2.8, 2.8, 3.6.1)

[112] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, September 2012. (Cited in 1.1, 5.2.3)

[113] Carlos A Melo, Jarno Drost, Patrick J Wijchers, Harmen van de Werken, Elzo de Wit, Joachim A F Oude Vrielink, Ran Elkon, Sónia A Melo, Nicolas Léveillé, Raghu Kalluri, Wouter de Laat, and Reuven Agami. eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. *Mol. Cell*, 49(3):524–535, February 2013. (Cited in 5.2.3)

[114] Marco A Mendoza-Parra, Mannu Walia, Martial Sankar, and Hinrich Gronemeyer. Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. *Mol. Syst. Biol.*, 7:538, 2011. (Cited in 4.1.1)

[115] Naoko Minegishi, Sei Morita, Masayoshi Minegishi, Shigeru Tsuchiya, Tasuke Konno, Norio Hayashi, and Masayuki Yamamoto. Expression of GATA transcription factors in myelogenous and lymphoblastic leukemia cells. *Int. J. Hematol.*, 65(3):239–249, April 1997. (Cited in 3.6.4)

[116] modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kherad-

pour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher A Bristow, Lijia Ma, Michael F Lin, Stefan Washietl, Bradley I Arshinoff, Ferhat Ay, Patrick E Meyer, Nicolas Robine, Nicole L Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D Brown, Rogerio Candeias, Joseph W Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y Tolstorukov, Sebastian Will, Artyom A Alekseyenko, Carlo Artieri, Benjamin W Booth, Angela N Brooks, Qi Dai, Carrie A Davis, Michael O Duff, Xin Feng, Andrey A Gorchakov, Tingting Gu, Jorja G Henikoff, Philipp Kapranov, Renhua Li, Heather K MacAlpine, John Malone, Aki Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K Powell, Nicole C Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E Sandler, Yuri B Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E Brenner, Michael R Brent, Lucy Cherbas, Sarah C R Elgin, Thomas R Gingeras, Robert Grossman, Roger A Hoskins, Thomas C Kaufman, William Kent, Mitzi I Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W Posakony, Bing Ren, Steven Russell, Peter Cherbas, Brenton R Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J Park, Susan E Celniker, Steven Henikoff, Gary H Karpen, Eric C Lai, David M MacAlpine, Lincoln D Stein, Kevin P White, and Manolis Kellis. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012):1787–1797, December 2010. (Cited in 4.1.1)

[117] Quaid Morris, Martha L Bulyk, and Timothy R Hughes. Jury remains out on simple models of transcription factor specificity. *Nat. Biotechnol.*, 29(6):483–484, June 2011. (Cited in 5.2.1)

[118] Mouse ENCODE Consortium, John A Stamatoyannopoulos, Michael Snyder, Ross Hardison, Bing Ren, Thomas Gingeras, David M Gilbert, Mark Groudine, Michael Bender, Rajinder Kaul, Theresa Canfield, Erica Giste, Audra Johnson, Mia Zhang, Gayathri Balasundaram, Rachel Byron, Vaughan Roach, Peter J Sabo, Richard Sandstrom, A Sandra Stehling, Robert E Thurman, Sherman M Weissman, Philip Cayting, Manoj Hariharan, Jin Lian, Yong Cheng, Stephen G Landt, Zhihai Ma, Barbara J Wold, Job Dekker, Gregory E Crawford, Cheryl A Keller, Weisheng Wu, Christopher Morrissey, Swathi A Kumar, Tejaswini Mishra, Deepti Jain, Marta Byrska-Bishop, Daniel Blankenberg, Bryan R Lajoie1, Gaurav Jain, Amartya Sanyal, Kaun-Bei Chen, Olgert Denas, James Taylor,

Gerd A Blobel, Mitchell J Weiss, Max Pimkin, Wulan Deng, Georgi K Marinov, Brian A Williams, Katherine I Fisher-Aylor, Gilberto DeSalvo, Anthony Kiralusha, Diane Trout, Henry Amrhein, Ali Mortazavi, Lee Edsall, David McCleary, Samantha Kuan, Yin Shen, Feng Yue, Zhen Ye, Carrie A Davis, Chris Zaleski, Sonali Jha, Chenghai Xue, Alex Dobin, Wei Lin, Meagan Fastuca, Huaien Wang, Roderic Guigó, Sarah Djebali, Julien Lagarde, Tyrone Ryba, Takayo Sasaki, Venkat S Malladi, Melissa S Cline, Vanessa M Kirkup, Katrina Learned, Kate R Rosenbloom, W James Kent, Elise A Feingold, Peter J Good, Michael Pazin, Rebecca F Lowdon, and Leslie B Adams. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, 13(8):418, August 2012. (Cited in 3.1, 3.3.1)

[119] Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young, and Martha L Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, 36(12):1331–1339, December 2004. (Cited in 1.2.2, 3.1)

[120] Patricia A J Muller and Karen H Vousden. p53 mutations in cancer. *Nat. Cell Biol.*, 15(1):2–8, 2013. (Cited in 1.1)

[121] Mariana Nacht, Tatiana Dracheva, Yuhong Gao, Takeshi Fujii, Yidong Chen, Audrey Player, Viatcheslav Akmaev, Brian Cook, Michael Dufault, Mindy Zhang, Wen Zhang, Mingzhou Guo, John Curran, Sean Han, David Sidransky, Kenneth Buetow, Stephen L Madden, and Jin Jen. Molecular characteristics of non-small cell lung cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 98(26):15203–15208, December 2001. (Cited in 1.1)

[122] Magdalene Nakou, Nicholas Knowlton, Mark B Frank, George Bertsias, Jeanette Osban, Clayton E Sandel, Helen Papadaki, Amalia Raptopoulou, Prodromos Sidiropoulos, Iraklis Kritikos, Ioannis Tassiulas, Michael Centola, and Dimitrios T Boumpas. Gene expression in systemic lupus erythematosus: bone marrow analysis differentiates active from inactive disease and reveals apoptosis and granulopoiesis signatures. *Arthritis Rheum.*, 58(11):3541–3549, November 2008. (Cited in 1.1)

[123] Jennifer L Nemhauser, Fangxin Hong, and Joanne Chory. Different plant hormones regulate similar

processes through largely nonoverlapping transcriptional responses. *Cell*, 126(3):467–475, August 2006. (Cited in 4.3.10)

[124] Shane Neph, Andrew B Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A Stamatoyannopoulos. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–1286, September 2012. (Cited in 3.1, 3.4.2, 3.6.3, 3.7, 5.2.3)

[125] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, Matthew T Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shinny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R Scott Hansen, Tanya Kutyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J MacCoss, Joshua M Akey, M A Bender, Mark Groudine, Rajinder Kaul, and John A Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, September 2012. (Cited in 3.7)

[126] Daniel E Newburger and Martha L Bulyk. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, 37(Database issue):D77–82, January 2009. (Cited in 3.1, 3.2.5)

[127] Li Ni, Can Bruce, Christopher Hart, Justine Leigh-Bell, Daniel Gelperin, Lara Umansky, Mark B Gerstein, and Michael Snyder. Dynamic and complex transcription factor binding during an inducible response in yeast. *Genes Dev.*, 23(11):1351–1363, June 2009. (Cited in 4.1.1, 4.5)

[128] Duncan T Odom, Nora Zizlsperger, D Benjamin Gordon, George W Bell, Nicola J Rinaldi, Heather L Murray, Tom L Volkert, Jorg Schreiber, P Alexander Rolfe, David K Gifford, Ernest Fraenkel, Graeme I Bell, and Richard A Young. Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662):1378–1381, February 2004. (Cited in 3.6.5)

[129] Yaron Orenstein, Chaim Linhart, and Ron Shamir. Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data. *PLoS ONE*, 7(9):e46145, 2012. (Cited in 3.1)

[130] Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. ChIP-Seq of transcription factors predicts

absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 106(51):21521–21526, December 2009. (Cited in 5.2.3)

[131] Jessica Pamment, Eleanor Ramsay, Michael Kelleher, David Dornan, and Kathryn L Ball. Regulation of the IRF-1 tumour modifier during the response to genotoxic stress involves an ATM-dependent signalling pathway. *Oncogene*, 21(51):7776–7785, November 2002. (Cited in 2.8)

[132] Xin Pan, Jie Zhao, Wei-Na Zhang, Hui-Yan Li, Rui Mu, Tao Zhou, Hai-Ying Zhang, Wei-Li Gong, Ming Yu, Jiang-Hong Man, Pei-Jing Zhang, Ai-Ling Li, and Xue-Min Zhang. Induction of SOX4 by DNA damage is critical for p53 stabilization and function. *Proc. Natl. Acad. Sci. U.S.A.*, 106 (10):3788–3793, March 2009. (Cited in 2.8)

[133] Pier Paolo Pandolfi, Matthew E Roth, Alar Karis, Mark W Leonard, E Dzierzak, Frank G Grosveld, James Douglas Engel, and Michael H Lindenbaum. Targeted disruption of the GATA3 gene causes severe abnormalities in the nervous system and in fetal liver haematopoiesis. *Nat. Genet.*, 11(1): 40–44, September 1995. (Cited in 3.6.4)

[134] Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10 (10):669–680, October 2009. (Cited in 1.2.4, 3.1)

[135] Mary Elizabeth Patti, Atul J Butte, Sarah Crunkhorn, Kenneth Cusi, Rachele Berria, Sangeeta Kashyap, Yoshinori Miyazaki, Isaac Kohane, Maura Costello, Robert Saccone, Edwin J Landaker, Allison B Goldfine, Edward Mun, Ralph DeFronzo, Jean Finlayson, C Ronald Kahn, and Lawrence J Mandarino. Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. *Proc. Natl. Acad. Sci. U.S.A.*, 100(14):8466–8471, July 2003. (Cited in 1.1)

[136] Joseph C Pearson, Derek Lemons, and William McGinnis. Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.*, 6(12):893–904, December 2005. (Cited in 1.1)

[137] A C Pease, D Solas, E J Sullivan, M T Cronin, C P Holmes, and S P Fodor. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 91(11): 5022–5026, May 1994. (Cited in 1.2.1)

[138] Zhi-yu Peng, Xin Zhou, Linchuan Li, Xiangchun Yu, Hongjiang Li, Zhiqiang Jiang, Guangyu Cao,

Mingyi Bai, Xingchun Wang, Caifu Jiang, Haibin Lu, Xianhui Hou, Lijia Qu, Zhiyong Wang, Jianru Zuo, Xiangdong Fu, Zhen Su, Songgang Li, and Hongwei Guo. Arabidopsis Hormone Database: a comprehensive genetic and phenotypic information database for plant hormone research in Arabidopsis. *Nucleic Acids Res.*, 37(Database issue):D975–82, January 2009. (Cited in 4.3.10)

[139] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, March 2011. (Cited in 3.1, 3.3.1, 3.7)

[140] J R Pollack, C M Perou, A A Alizadeh, M B Eisen, A Pergamenschikov, C F Williams, S S Jeffrey, D Botstein, and P O Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, 23(1):41–46, September 1999. (Cited in 1.2.1)

[141] M L Quek, D I Quinn, S Daneshmand, and J P Stein. Molecular prognostication in bladder cancer–a current perspective. *Eur. J. Cancer*, 39(11):1501–1510, July 2003. (Cited in 1.1)

[142] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, Piero Carninci, Carsten O Daub, Alistair R R Forrest, Julian Gough, Sean Grimmond, Jung-Hoon Han, Takehiro Hashimoto, Winston Hide, Oliver Hofmann, Atanas Kamburov, Mandeep Kaur, Hideya Kawaji, Atsutaka Kubosaki, Timo Lassmann, Erik van Nimwegen, Cameron Ross MacPherson, Chihiro Ogawa, Aleksandar Radovanovic, Ariel Schwartz, Rohan D Teasdale, Jesper Tegnér, Boris Lenhard, Sarah A Teichmann, Takahiro Arakawa, Noriko Ninomiya, Kayoko Murakami, Michihira Tagami, Shiro Fukuda, Kengo Imamura, Chikatoshi Kai, Ryoko Ishihara, Yayoi Kitazume, Jun Kawai, David A Hume, Trey Ideker, and Yoshihide Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, March 2010. (Cited in 3.6.4)

[143] Emma Redhead and Timothy L Bailey. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, 8:385, 2007. (Cited in 2.1, 2.2.2, 2.3, 2.4.1, 2.6)

[144] Bing Ren, François Robert, John J Wyrick, Oscar Aparicio, Ezra G Jennings, Itamar Simon, Julia Zeitlinger, Jorg Schreiber, Nancy Hannett, Elenita Kanin, Thomas L Volkert, Christopher J Wilson, Stephen P Bell, and Richard A Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, December 2000. (Cited in 3.1)

[145] Guillaume Rey, François Cesbron, Jacques Rougemont, Hans Reinke, Michael Brunner, and Felix Naef. Genome-wide and phase-specific DNA-binding rhythms of BMAL1 control circadian output functions in mouse liver. *PLoS Biol.*, 9(2):e1000595, February 2011. (Cited in 4.5)

[146] Daniel R Rhodes, Jianjun Yu, K Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pandey, and Arul M Chinnaiyan. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U.S.A.*, 101(25):9309–9314, June 2004. (Cited in 2.8)

[147] Todd Riley, Eduardo Sontag, Patricia Chen, and Arnold Levine. Transcriptional control of human p53-regulated genes. *Nat. Rev. Mol. Cell Biol.*, 9(5):402–412, May 2008. (Cited in 2.8)

[148] Anna Ritz, Gregory Shakhnarovich, Arthur R Salomon, and Benjamin J Raphael. Discovery of phosphorylation motif mixtures in phosphoproteomics data. *Bioinformatics*, 25(1):14–21, January 2009. (Cited in 2.1)

[149] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651–657, August 2007. (Cited in 2.1)

[150] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77, 2011. (Cited in 3.6.2, 3.2)

[151] Frederick P Roth, Jason D Hughes, Preston W Estep, and George M Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, 16(10):939–945, October 1998. (Cited in 2.1)

[152] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, 27(1):66–75, January 2009. (Cited in 4.3.7)

[153] Laurent Sansregret and Alain Nepveu. The multiple roles of CUX1: insights from mouse models and cell-based assays. *Gene*, 412(1-2):84–94, April 2008. (Cited in 3.6.5)

[154] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, October 1995. (Cited in 1.2.1, 4.1.1)

[155] Dominic Schmidt, Michael D Wilson, Benoit Ballester, Petra C Schwalie, Gordon D Brown, Aileen Marshall, Claudia Kutter, Stephen Watt, Celia P Martinez-Jimenez, Sarah Mackay, Iannis Talianidis, Paul Flicek, and Duncan T Odom. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, May 2010. (Cited in 3.1, 3.7, 5.2.4)

[156] T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, October 1990. (Cited in 1.2.5)

[157] Jean-Jacques Schott, D Woodrow Benson, Craig T Basson, William Pease, G Michael Silberbach, Jeffrey P Moak, Barry J Maron, Christine E Seidman, and J G Seidman. Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science*, 1998. (Cited in 1.1)

[158] Marcel H Schulz, William E Devanny, Anthony Gitter, Shan Zhong, Jason Ernst, and Ziv Bar-Joseph. DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst Biol*, 6:104, 2012. (Cited in 1.3, 3.7, 4.1.1, 4.5)

[159] A Schulze and J Downward. Navigating gene expression using microarrays–a technology review. *Nat. Cell Biol.*, 3(8):E190–5, August 2001. (Cited in 1.2.1)

[160] Daniel Schwartz and George M Church. Collection and motif-based prediction of phosphorylation sites in human viruses. *Sci Signal*, 3(137):rs2, 2010. (Cited in 2.1)

[161] Brock L Schweitzer, Kelly J Huang, Meghana B Kamath, Alexander V Emelyanov, Barbara K Birshtein, and Rodney P DeKoter. Spi-C has opposing effects to PU.1 on gene expression in

progenitor B cells. *J. Immunol.*, 177(4):2195–2207, August 2006. (Cited in 3.3)

[162] Helena Shaked, Idit Shiff, Miriam Kott-Gutkowski, Zahava Siegfried, Ygal Haupt, and Itamar Simon. Chromatin immunoprecipitation-on-chip reveals stress-dependent p53 occupancy in primary normal cells but not in established cell lines. *Cancer Res.*, 68(23):9671–9677, December 2008. (Cited in 2.5.1)

[163] Eilon Sharon, Shai Lubliner, and Eran Segal. A feature-based approach to modeling protein-DNA interactions. *PLoS Comp Biol*, 4(8):e1000154, 2008. (Cited in 1.2.5, 5.2.1)

[164] Xiaohai Shi, D V Bosenko, N S Zinkevich, S Foley, D R Hyde, E V Semina, and Thomas S Vihtelic. Zebrafish pitx3 is necessary for normal lens and retinal development. *Mech. Dev.*, 122(4):513–527, April 2005. (Cited in 3.3)

[165] David Q Shih, Markus Bussen, Ephraim Sehayek, Meenakshisundaram Ananthanarayanan, Benjamin L Shneider, Frederick J Suchy, Sarah Shefer, Jaya S Bollileni, Frank J Gonzalez, Jan L Breslow, and Markus Stoffel. Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism. *Nat. Genet.*, 27(4):375–382, April 2001. (Cited in 3.6.5)

[166] Rahul Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS ONE*, 5(3):e9722, 2010. (Cited in 1.2.5, 5.2.1)

[167] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8): 1034–1050, August 2005. (Cited in 3.3.2, 3.6.3)

[168] Saurabh Sinha. Discriminative motifs. *J. Comput. Biol.*, 10(3-4):599–615, 2003. (Cited in 2.1, 5.2.2)

[169] Saurabh Sinha. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, 22(14):e454–63, July 2006. (Cited in 2.1, 2.2.3, 2.3, 2.3, 2.3.3, 2.3.3, 2.3.4, 2.4.1, 2.6, 2.6.5)

[170] Saurabh Sinha and Martin Tompa. YMF: A program for discovery of novel transcription factor

binding sites by statistical overrepresentation. *Nucleic Acids Res.*, 31(13):3586–3588, July 2003. (Cited in 2.1)

[171] Julia Sivriver, Naomi Habib, and Nir Friedman. An integrative clustering and modeling algorithm for dynamical gene expression data. *Bioinformatics*, 27(13):i392–400, July 2011. (Cited in 4.1.1)

[172] Andrew D Smith, Pavel Sumazin, and Michael Q Zhang. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl. Acad. Sci. U.S.A.*, 102(5):1560–1565, February 2005. (Cited in 2.1, 2.2.1, 2.3, 2.4.1, 2.4.1, 2.6, 5.2.2)

[173] Alex Yick-Lun So, Samantha B Cooper, Brian J Feldman, Mitra Manuchehri, and Keith R Yamamoto. Conservation analysis predicts in vivo occupancy of glucocorticoid receptor-binding sequences at glucocorticoid-induced genes. *Proc. Natl. Acad. Sci. U.S.A.*, 105(15):5745–5749, April 2008. (Cited in 3.3, 3.6.3)

[174] Raymond E Soccio, Geetu Tuteja, Logan J Everett, Zhaoyu Li, Mitchell A Lazar, and Klaus H Kaestner. Species-specific strategies underlying conserved functions of metabolic transcription factors. *Mol. Endocrinol.*, 25(4):694–706, April 2011. (Cited in 3.1)

[175] Kristen M Sokalski, Stephen K H Li, Ian Welch, Heather-Anne T Cadieux-Pitre, Marek R Gruca, and Rodney P DeKoter. Deletion of genes encoding PU.1 and Spi-B in B cells impairs differentiation and induces pre-B cell acute lymphoblastic leukemia. *Blood*, 118(10):2801–2808, September 2011. (Cited in 3.3)

[176] Le Song, Mladen Kolar, and Eric P Xing. KELLER: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–36, June 2009. (Cited in 4.1.1)

[177] François Spitz and Eileen E M Furlong. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626, September 2012. (Cited in 3.1)

[178] Marc Sultan, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, August 2008. (Cited in 4.1.1)

[179] Mikita Suyama, Eoghan D Harrington, Svetlana Vinokourova, Magnus von Knebel Doeberitz, Osamu Ohara, and Peer Bork. A network of conserved co-occurring motifs for the regulation of alternative splicing. *Nucleic Acids Res.*, 38(22):7916–7926, December 2010. (Cited in 2.1)

[180] Pamela J Swiatek and Thomas Gridley. Perinatal lethality and defects in hindbrain development in mice homozygous for a targeted mutation of the zinc finger gene Krox20. *Genes Dev.*, 7(11): 2071–2084, November 1993. (Cited in D.1)

[181] Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre Rouzé, and Yves Moreau. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, 9(2):447–464, 2002. (Cited in 2.1)

[182] Michael D Thompson and Satdarshan P S Monga. WNT/beta-catenin signaling in liver health and disease. *Hepatology*, 45(5):1298–1305, May 2007. (Cited in 3.3)

[183] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012. (Cited in 1.2.6)

[184] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Régnier, Nicolas Si-

monis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23(1):137–144, January 2005. (Cited in 2.4.3, 2.4.3, 2.7.2, 2.10)

[185] Axel Visel, Edward M Rubin, and Len A Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, September 2009. (Cited in 1.2)

[186] Yu-Jui Yvonne Wan, Dahsing An, Yan Cai, Joyce J Repa, Tim Hung-Po Chen, Monica Flores, Catherine Postic, Mark A Magnuson, Ju Chen, Kenneth R Chien, Samuel French, David J Mangelsdorf, and Henry M Sucov. Hepatocyte-specific mutation establishes retinoid X receptor alpha as a heterodimeric integrator of multiple physiological processes in the liver. *Mol. Cell. Biol.*, 20 (12):4436–4444, June 2000. (Cited in 3.3)

[187] D G Wang, J B Fan, C J Siao, A Berno, P Young, R Sapolsky, G Ghandour, N Perkins, E Winchester, J Spencer, L Kruglyak, L Stein, L Hsie, T Topaloglou, E Hubbell, E Robinson, M Mittmann, M S Morris, N Shen, D Kilburn, J Rioux, C Nusbaum, S Rozen, T J Hudson, R Lipshutz, M Chee, and E S Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, May 1998. (Cited in 1.2.1)

[188] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009. (Cited in 1.2.3)

[189] Chia-Lin Wei, Qiang Wu, Vinsensius B Vega, Kuo Ping Chiu, Patrick Ng, Tao Zhang, Atif Shahab, How Choong Yong, Yutao Fu, Zhiping Weng, JianJun Liu, Xiao Dong Zhao, Joon-Lin Chew, Yen Ling Lee, Vladimir A Kuznetsov, Wing-Kin Sung, Lance D Miller, Bing Lim, Edison T Liu, Qiang Yu, Huck-Hui Ng, and Yijun Ruan. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124(1):207–219, January 2006. (Cited in 2.8)

[190] Bartek Wilczyński and Norbert Dojer. BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics*, 25(2):286–287, January 2009. (Cited in 4.1.1)

[191] Bartek Wilczyński and Eileen E M Furlong. Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.*, 2010. (Cited in 4.1.1)

[192] Christian Wolfrum, Esra Asilmaz, Edlira Luca, Jeffrey M Friedman, and Markus Stoffel. Foxa2 regulates lipid metabolism and ketogenesis in the liver during fasting and in diabetes. *Nature*, 432 (7020):1027–1032, December 2004. (Cited in 3.6.5)

[193] Kyoung-Jae Won, Bing Ren, and Wei Wang. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, 11(1):R7, 2010. (Cited in 3.7)

[194] William K K Wu, Chi H Cho, Chung W Lee, Daiming Fan, Kaichun Wu, Jun Yu, and Joseph J Y Sung. Dysregulation of cellular signaling in gastric cancer. *Cancer Lett.*, 295(2):144–153, September 2010. (Cited in 1.1)

[195] Xiaohui Xie, Jun Lu, E J Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, March 2005. (Cited in 3.3, 3.6.3, 5.2.4)

[196] Eric P Xing, Wei Wu, Michael I Jordan, and Richard M Karp. LOGOS: a modular Bayesian model for de novo motif detection. *Proc IEEE Comput Soc Bioinform Conf*, 2:266–276, 2003. (Cited in 2.1)

[197] Junzhe Xu, Jingchun Sun, Jingchun Chen, Lily Wang, Anna Li, Matthew Helm, Steven L Dubovsky, Silviu-Alin Bacanu, Zhongming Zhao, and Xiangning Chen. RNA-Seq analysis implicates dysregulation of the immune system in schizophrenia. *BMC Genomics*, 13 Suppl 8:S2, 2012. (Cited in 1.1)

[198] Alper Yilmaz, Maria Katherine Mejia-Guerra, Kyle Kurz, Xiaoyu Liang, Lonnie Welch, and Erich Grotewold. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res.*, 39(Database issue):D1118–22, January 2011. (Cited in 4.3.9)

[199] Ming Yu, Laura Riva, Huafeng Xie, Yocheved Schindler, Tyler B Moran, Yong Cheng, Duonan Yu, Ross Hardison, Mitchell J Weiss, Stuart H Orkin, Bradley E Bernstein, Ernest Fraenkel, and Alan B Cantor. Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell*, 36(4):682–695, November 2009. (Cited in 2.1)

[200] Julia Zeitlinger, Robert P Zinzen, Alexander Stark, Manolis Kellis, Hailan Zhang, Richard A

Young, and Michael Levine. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev.*, 21(4): 385–390, February 2007. (Cited in 3.1)

[201] Luwen Zhang, Xiangchun Ju, Yumin Cheng, Xiuyun Guo, and Tieqiao Wen. Identifying Tmem59 related gene regulatory network of mouse neural stem cell from a compendium of expression profiles. *BMC Syst Biol*, 5:152, 2011. (Cited in 3.6.4)

[202] Pili Zhang, Myriam Bennoun, Cécile Gogard, Pascale Bossard, Isabelle Leclerc, Axel Kahn, and Mireille Vasseur-Cognet. Expression of COUP-TFII in metabolic tissues during development. *Mech. Dev.*, 119(1):109–114, November 2002. (Cited in 3.3)

[203] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008. (Cited in 4.3.7)

[204] Yue Zhao and Gary D Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, 29(6):480–483, June 2011. (Cited in 1.2.2, 3.1, 3.2.1, 3.2.1, 3.2.1, 3.2.2, 3.2.2, 2, 3.6.1, 3.6.1, 3.6.2, 3.7, 5.2.1)

[205] Shan Zhong, Xin He, and Ziv Bar-Joseph. Predicting tissue-specific transcription factor binding sites. *Submitted*. (Cited in 1.3, 5.2.3)

[206] Qing Zhou and Jun S Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916, April 2004. (Cited in 1.2.5, 5.2.1)

[207] Cong Zhu, Kelsey J R P Byers, Rachel Patton McCord, Zhenwei Shi, Michael F Berger, Daniel E Newburger, Katrina Saulrieta, Zachary Smith, Mita V Shah, Mathangi Radhakrishnan, Anthony A Philippakis, Yanhui Hu, Federico De Masi, Marcin Pacek, Andreas Rolfs, Tal Murthy, Joshua Labaer, and Martha L Bulyk. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research*, 19(4):556–566, April 2009. (Cited in 1.2.2, 1.2.2, 3.1, 4, 3.6.2, 3.7)

[208] Robert P Zinzen, Charles Girardot, Julien Gagneur, Martina Braun, and Eileen E M Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70,

November 2009. (Cited in 4.5)