Computational Regulatory Genomics: Motifs, Networks, and Dynamics

by

Pouya Kheradpour

B.S. Computer Science, University of Illinois at Urbana-Champaign (2005)

M.S. Computer Science, University of Illinois at Urbana-Champaign (2005)

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

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2012

© Pouya Kheradpour, MMXII. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly paper and electronic copies of this thesis document in whole or in part.

Author . . . .
Department of Electrical Engineering and Computer Science
February 3, 2012

Certified by . . . .
Manolis Kellis
Associate Professor of Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Chair of the Committee on Graduate Students

Computational Regulatory Genomics: Motifs, Networks, and Dynamics

by

Pouya Kheradpour

Submitted to the Department of Electrical Engineering and Computer Science
on February 3, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

Gene regulation, the process responsible for taking a static genome and producing the diversity and complexity of life, is largely mediated through the sequence specific binding of regulators. The short, degenerate nature of the recognized elements and the unknown rules through which they interact makes deciphering gene regulation a significant challenge.

In this thesis, we utilize comparative genomics and other approaches to exploit large-scale experimental datasets and better understand the sequence elements and regulators responsible for regulatory programs. In particular, we develop new computational approaches to (1) predict the binding sites of regulators using the genomes of many, closely related species; (2) understand the sequence motifs associated with transcription factors; (3) discover and characterize microRNAs, an important class of regulators; (4) use static predictions for binding sites in conjunction with chromatin modifications to better understand the dynamics of regulation; and (5) systematically validate the predicted motif instances using a massively parallel reporter assay.

We find that the predictions made by our algorithms are of high quality and are comparable to those made by leading experimental approaches. Moreover, we find that experimental and computational approaches are often complementary. Regions experimentally identified to be bound by a factor can be species and cell line specific, but they lack the resolution and unbiased nature of our predictions. Experimentally identified miRNAs have unmistakable signs of being processed, but cannot provide the same insights our machine learning framework does. Further emphasizing the importance of integration, combining chromatin mark annotations and gene expression from multiple cell types with our static motif instances allows for increasing our power and making additional biologically relevant insights.

We successfully apply the algorithms in this thesis to 29 mammals and 12 flies and expect them to be applicable to other clades of eukaryotic species. Moreover, we find that our performance has not yet plateaued and believe these methods will continue to be relevant as sequencing becomes increasingly commonplace and thousands of genomes become available.

Thesis Supervisor: Manolis Kellis
Title: Associate Professor of Computer Science

Acknowledgments

First and foremost, I am deeply indebted to my supervisor Manolis Kellis for his guidance throughout my doctoral work and his consistent encouragement as I explored various aspects of regulatory genomics. Much of this thesis would never had completed had it not been for his sustained enthusiasm and direction. I would also like to thank David Gifford and Martha Bulyk for graciously agreeing to be my committee members and for their advice toward completing my dissertation and regarding my future career.

Science frequently benefits from collaboration and this work is no exception; many members of the Kellis Lab have undeniably influenced this thesis, particularly Alexander Stark, Leopold Parts, and Jason Ernst. Further, many of the results presented within this thesis were possible only because of data and experimental work from the labs of Bradley Bernstein, Tarjei Mikkelsen, Kerstin Lindblad-Toh and the ENCODE, modENCODE, 12 fly, and 29 mammal consortia. During the past six and a half years I have had numerous comments and suggestions during my presentations and as part of these consortia, much of which positively influenced this thesis and for that I am grateful.

People require social interaction in order to maintain happiness and sanity, a fact that has been particularly true for me during my longest and most unstructured stage of development: graduate school. For providing me with this necessity, I would like to thank Seb Neumayer, Matt Rasmussen, Chris Evans, John Kelleher, Mike Lin, and especially Tracy Tat.

I have been truly fortunate to have two aunts, Khalehs Simin and Monir, and their families in the Boston area. They have opened their homes to me and kept me well nourished on countless occasions. Finally, I have to thank my parents, Albert and Jacklin, and my siblings Saba and Nima, for their unconditional love and continued support in pursing my interests.

Contents

Chapter 1

Introduction

1.1 Motivations

While gene regulation is vital to all life, our understanding of the underlying players and their precise roles remains incomplete. This thesis presents the design and application of algorithms seeking to increase our knowledge of this basic process. Interest in computational regulatory genomics has greatly increased in recent years due to rapid advances experimental techniques, particularly the exponential drop in the cost of sequencing.

The rapid rate of this advancement is apparent when examining publications in the field. In 2000, the first draft of the *Drosophila* genome was published (∼120 megabases; Adams et al., 2000). A year later a decade long project culminated with the publication of the first draft of the human genome (∼3 gigabases; Lander et al., 2001; Venter et al., 2001). Today we have the sequence of 12 flies (Stark et al., 2007b; *Drosophila* 12 Genomes Consortium, 2007), dozens of mammals (Lindblad-Toh et al., 2011), and hundreds of additional eukaryotes (Kersey et al., 2009). Further, sequencing is now applied to even individuals of a species with dozens already sequenced and hundreds planned (Kaiser, 2008).

We benefit from this new abundance of data throughout this thesis — from predicting regulators, the patterns they recognize, to their specific instances in the genome. We also use these technologies to predict regulators and test them using a high throughput enhancer assay. This dramatic rate of technological advancement has made it difficult to predict the potential scope and direction of research making it an exciting time to be involved in computational biology.

1.2 Relevant biology background

This section will provide an overview of the biological concepts necessary to understand the work in this thesis. The interested reader is encouraged to consult a more thorough treatment of the relevant topics, which is available in text books on molecular biology (Alberts et al., 2002; Watson et al., 2007; Lodish et al., 2007), gene regulation (Latchman, 2010), and computational biology (Durbin et al., 1998; Jones and Pevzner, 2004) and through online resources (e.g., Wikipedia). This treatment will also ignore most exceptions (which exist for nearly every biological statement) unless they are necessary for understanding a concept in this thesis.

1.2.1 Basics of molecular biology

All life on earth is made up of building blocks called cells. Genetic material in the form of deoxyribonucleic acid (DNA) is found in nearly all of these cells and encodes the primary "blueprints" for the development and response of the cells to external stimuli. This thesis will focus on eukaryotes, a class of organisms which includes fungi, plants and animals and for which most of the DNA is found in the nucleus. The language of DNA has a 4 character system of adenine (A), cytosine (C), guanine (G) and thymine (T), which are referred to as bases (bp) or nucleotides (nt). DNA is structured as a double-helix with complementary base-pairing (A to T, G to C) to facilitate easy replication, a feature noted since its initial characterization (Watson and Crick, 1953).

DNA is organized into chromosomes which are essentially long strings of the 4 bases. Obtaining these strings is the desired result of sequencing, but sequencing errors and repetitive regions can complicate the complete recovery of each

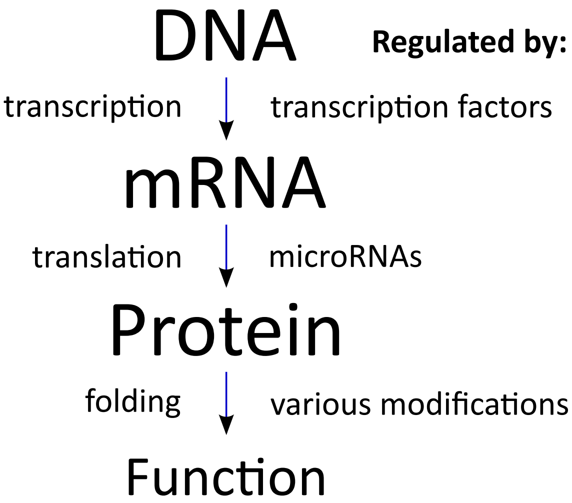


Figure 1-1: Central dogma of molecular biology and an incomplete list of the major regulators involved.

chromosome's sequence. Chromosomes contain specific substrings that encode functional elements. These functions can overlap with the same base having multiple roles. Finding the coordinates of specific classes of functional elements will be one of the primary focuses of this thesis.

Large areas of the genome (from tens to as many as millions of bases) are transcribed into ribonucleic acid (RNA) by RNA polymerase proteins and are referred to as genes. The direction of RNA synthesis (like DNA synthesis) is always the same: from $5'$ to $3'$ (these refer to specific atoms in the chemical backbone of DNA and RNA) and starts from the transcription start site (TSS) and continues until the transcription end site (TES; also known as the poly(A) site). DNA/RNA sequences in this paper will always be indicated in this order. Compared to DNA, RNA has specific chemical differences, including the nucleotide uracil (U) instead of T and does not always exist in a double-stranded form. RNAs take on a number of roles in the cell and for some viruses can even be the primary carrier of genetic information.

Messenger RNAs (mRNAs) and are one of the primary types of RNAs encoded by the genome. They are transcribed (or expressed) and then are processed ('spliced') removing sections referred to as introns leaving the exons. These ma-

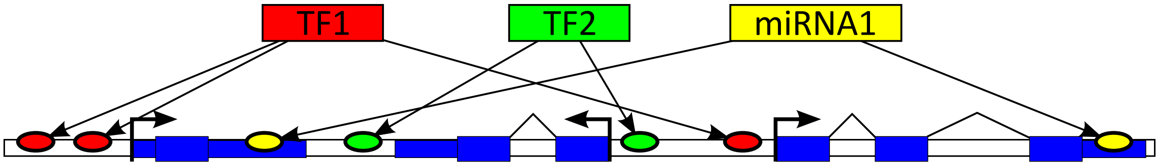


Figure 1-2: A simple model of gene regulation. Regulators (top) each have an associated motif to which they bind. Transcription factors (TFs) bind near the TSSs of genes they regulate, while microRNAs (miRNAs) bind to the 3′ UTRs of mRNAs.

ture transcripts are then translated into proteins using a code based on sliding non-overlapping windows of length 3. Each of the 64 3-base sequences (or codons) specifies one of 20 amino acids. Translation always begins with an AUG and ends with a UAA, UAG, or UGA. The portion of the processed mRNA that is untranslated before the coding portion is referred to as the 5′ UTR and the portion following it is the 3′ UTR. The transcription of DNA to RNA and the subsequent translation to proteins is referred to as the central dogma of molecular biology (Figure 1-1).

In eukaryotes chromosomes are wrapped around proteins called histones creating nucleosomes. Both the DNA and histones can undergo semi-stably inherited covalent changes referred to as epigenetic modifications. Experimental techniques, such as ChIP-chip or ChIP-seq (Figure 1-4), are used to read the modification state of each region of the genome (with resolution depending on the type of modification and the technique). Cataloging the modification state (or epigenome) is a substantial effort because many dozens of known modifications exist and must be annotated for each cell type (Bernstein et al., 2010; Celniker et al., 2009; Consortium, 2011b). These changes are correlated with various functional properties (Suzuki and Bird, 2008; Kouzarides, 2007; Ernst and Kellis, 2010), a fact that we will exploit in this thesis when predicting functional regions of the genome.

1.2.2 Gene regulation

While every cell in the body contains essentially the same DNA, cells themselves have dramatically different morphologies, behavior, and functions. This diversity is largely driven through the specific regulation of which genes are active in a cell. In turn, regulation occurs in every step between the DNA sequence and protein function (Figure 1-1). This section will briefly go through the mechanisms through which this occurs.

In vertebrates, some genes are responsive to DNA methylation near their upstream of their TSS. DNA methylation can effectively turn off a gene (Suzuki and Bird, 2008) and can be stably inherited across cell devisions. In mammals most cytosines (C) that are followed by a guanine (G) are methylated. Because of the specific chemistry involved, methylated cytosines have a propensity to be mutated to thymine (T). Consequently the genome as a whole is depleted of CG dinucleotides. CpG islands are particularly common near housekeeping genes where DNA is not methylated in the germline and consequently CGs are not depleted.

Transcription itself is a five step process: pre-initiation, initiation, clearance, elongation and termination. The primarily regulators involved in this process are proteins called transcription factors (TFs). Each TF recognizes a specific pattern to which it binds. This sequence, called a motif, can be of variable length (5-20+ bases) and can be degenerate (e.g., recognizing either an A or a G at a specific position). Further, TFs can bind a variable distance from the gene: while there is a clear enrichment of TF binding sites near the genes they target, there are also examples of distal binding sites, called enhancers, many thousands of bases away from a TSS. The various steps of transcription can be individually regulated as well: for example, polymerase can pause at a promoter and fail to produce full length transcripts (Core and Lis, 2008).

The product of transcription is a precursor mRNA (pre-mRNA). The subsequent splicing is regulated through the binding of factors called exonic/intronic splicing enhancers/silencers (reviewed in Wang and Burge, 2008). Alternative splicing, along with alternative transcription start and end sites, can lead different isoforms of the same gene. Current evidence suggests that more than two-thirds of human genes and two-fifths of *Drosophila* genes undergo some form of alternative splicing (Benjamin J., 2006), making it an important source of diversity for multicellular organisms. Although not explicitly discussed in this thesis, many of the approaches utilized here are also applicable to predict instances of splicing-related regulators.

After splicing the mature mRNA is exported from the nucleus into the cytoplasm where it is ready for translation. Several regulators including pumilio proteins (Wharton et al., 1998) and microRNAs (miRNAs) can bind to the RNA through sequence specificity and influence degradation or translation of the transcript.

MicroRNAs are one of the most important regulators (reviewed in Winter et al., 2009; Bartel, 2004) and are of particular interest to this thesis. MicroRNA genes, which can be over 1 kb in length, are transcribed from the genome and fold into $\sim$80-nt hairpin (Figure 1-3). These are then processed first in the nucleus and then in the cytoplasm to create a short ($\sim$22-nt) mature miRNA. These short RNAs bind to the $3'$ UTR and sometimes coding region (Stark et al., 2007b; Schnall-Levin et al., 2010) of a gene. Recognition of bases 2-7 of the miRNA (the seed) through canonical base pairing (A-U, G-C, not G-U) is largely responsible for the specificity of a miRNA in animals. Additional strength is added by a match to the 8th base, and having a match that is followed by an A and base pairing to the remaining portion of the miRNA (Lewis et al., 2005). Recognition by a miRNA leads to repression of the target mRNA effectively reducing the expression of a gene.

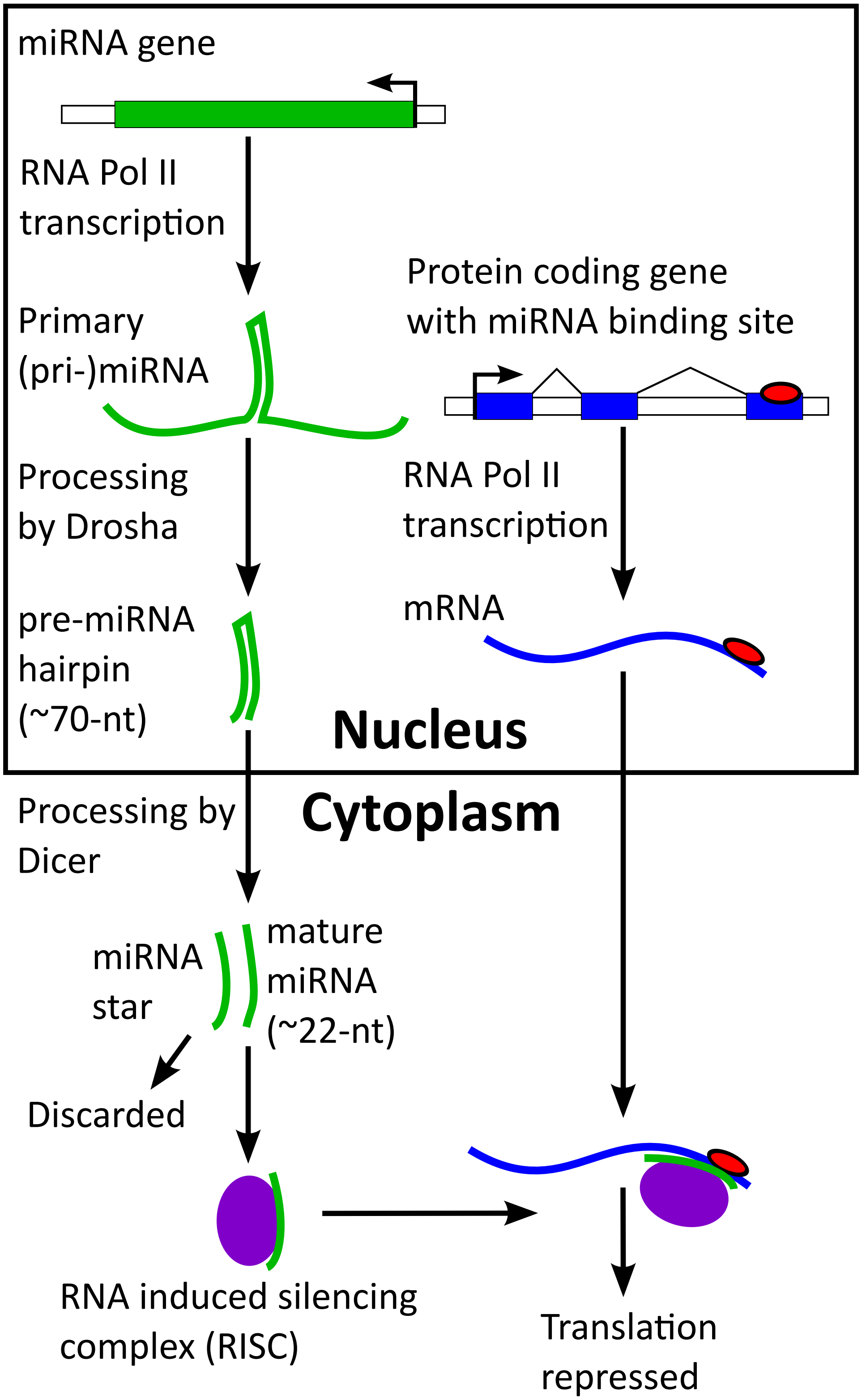


Figure 1-3: Biogenesis and function of miRNAs.

1.3 Comparative genomics

Species on earth can be placed into a tree or phylogeny indicating their relationships (incomplete lineage sorting and horizontal gene transfer make this only roughly true, however we will ignore it in this thesis as it is relatively rare in the species we consider). When species have experienced sufficiently little divergence, genomes can be aligned to predict orthologous bases and identify specific mutations. A number of procedures exist for this problem and whole genome alignment is an active area of research. We use alignments provided by the UCSC genome browser which are generated using MULTIZ (Blanchette et al., 2004).

Comparative genomics is the technique of taking advantage of this evolutionary history to better understand the genome. It can be difficult to predict even well specified genomic elements such as protein coding genes because we do not know all the rules that govern them. A popular and long successful strategy is essentially identify places that have a conspicuous reduction of mutations (Rubin et al., 2000; Kellis et al., 2003). Of course the cell cannot see the evolutionary history of a segment of DNA. However, a lack of mutations suggests that a segment is important and thus undergoing purifying selection. Beyond a lack of mutations, for some elements such as genes the specific pattern of mutations that do occur can be suggestive of this specific function (evaluated in Lin et al., 2008).

This thesis will make extensive use of comparative genomics to predict functional elements. As we will show, we have sufficient power now to discover regulatory motifs with the number of species available. However, in most cases we still do not have sufficient signal to find short, poorly specified motif instances with high specificity. As more genomes are sequenced, the specific mutational pattern can be examined to determine the likelihood of a position evolving like a motif match. Consequently, comparative algorithms will continue to be of interest and will have to be designed to deal with additional genomes.

1.4 Relevant experimental techniques

This section will give an overview the large-scale experimental techniques utilized in this thesis (with the exception of MPRA, which will be described in Chapter 6). Understanding these techniques and their potential drawbacks is important for interpreting the results of this thesis and understanding the conclusions drawn.

1.4.1 DNA Sequencing

DNA sequencing refers to the process of obtaining the order of bases in a DNA molecule — data that is absolutely necessary for essentially all the analysis found in this thesis. Substantial efforts over the past few decades have led to a dramatic reduction in the cost of sequencing (Schuster, 2008), which has had marked changes to how many molecular biology experiments are carried out.

Sanger sequencing, originally described in the mid-1970s (Sanger et al., 1977), was one of the first sequencing techniques and also the primary method used to sequence all the genomes utilized in this thesis. The basic protocol involves using labeled dideoxynucleotide triphosphates as DNA chain terminators. By running on a gel the product of the reaction of a sample with each of the four dideoxynucleotides (and the deoxynucleotides), one can read the bases in order by comparing the relative sizes of the resulting synthesized DNA. While the initial technique was a labor intensive process, alterations to the original chemistry and automation procedures led to the feasibility of sequencing large mammalian genomes (reviewed in Alterovitz et al., 2009).

Because this technology was only able to produce contiguous sequences of length no greater than a few hundred bases, additional techniques were necessary for decoding larger genomic regions (e.g., entire chromosomes). The whole-genome shotgun approach is presently the dominant strategy for this purpose and was first applied to an animal genome to sequence *Drosophila melanogaster* (Adams et al., 2000). This approach fragments the genome into smaller pieces which are selected at random for sequencing. Subsequent computational techniques are then used to assemble these fragments by utilizing their overlapping base pairs (Batzoglou et al., 2002; Huang et al., 2003). Because of this random selection procedure, regions of the genome can be missed by chance requiring a high coverage of sequencing in order to ensure each base is likely to be sequenced at least once.

Significant pressure to further reduce the price of sequencing led to the development of sequencing techniques that were highly parallelized and able to decode many more bases at a given cost (Shendure and Ji, 2008). Two commercial sequencing platforms based on a cyclic-array procedure (Shendure et al., 2005; Margulies et al., 2005) that produced data used in this thesis include 454 Genome Sequencers (Roche Applied Science) and the Illumina Genome Analyzer. These produce shorter contiguous segments, but with a vastly higher total amount of sequence. This made these technologies particularly suitable for sequencing miRNAs (used in Chapter 4) and for ChIP-seq (used in Chapters 2, 3, 5), although genome assembly using these short reads has also been investigated (Zerbino and Birney, 2008).

Sequencing and the subsequent assembly can suffer from a number of problems that can lead to missing or inaccurate sequence (Pop et al., 2002). The algorithms presented in this thesis, particularly that of Chapter 2, are designed to be robust against these errors by only minimally penalizing them. Moreover, many of the results presented in this thesis are statistical in nature and consequently are not strongly effected by the relatively few errors that occur due to sequencing.

1.4.2 mRNA expression analysis

While all cells have essentially the same DNA, a different complement of genes is expressed in each cell type. Consequently, determining the specific genes that are active in a given sample of cells is of great interest to understanding the underlying biology. While it is generally desired to identify the specific proteins that are active in a cell type, this is technically difficult (Garbis et al., 2005). However, assaying nucleic acids is considerably simpler and very frequently used as a surrogate for the protein levels.

One popular approach for measuring the level of mRNAs in a sample is the DNA microarray (Schena et al., 1995). Microarrays contain tens to millions spots each with many copies of a single stranded DNA probe. Because the location of these probes is known, by placing a sample of nucleic acid (usually DNA complementary to an mRNA sample) and identifying the level to which the spots hybridize to the sample, a quantitative measure of the amount of RNA can be obtained. DNA microarrays can be made either using cDNA probes corresponding to fragments of mRNAs (Cheung et al., 1999; Duggan et al., 1999; DeRisi et al., 1996) or using synthesized oligonucleotide arrays (Irizarry et al., 2003).

cDNA microarrays are generated from a library of mRNAs and have probes that can be hundreds of base pairs long. However, because it is difficult to deposit consistent amounts of probe in each spot, generally two samples are hybridized to each array, using a separate florescent label for each. By comparing the relative ratio of the two colors for each spot, an estimate of the abundance of each transcript across multiple samples can be made.

In contrast, oligonucleotide arrays have probes that are arbitrarily synthesized sequences of relatively short length (10-200 bp). Because these sequences may not uniquely identify a transcript and in order to reduce noise, typically many such probes are designed for each mRNA and the expression level of each transcript is estimated by combining these values (Irizarry et al., 2003). In Chapter 6, we exploit this synthesis procedure not for expression analysis but rather to produce a large library of arbitrary sequences.

Both DNA microarray technologies suffer from what can be very high noise levels (Reis-Filho et al., 2006). Moreover, due to the design assumptions, microarray manufacturers advise against the comparison of expression between genes and cross-hybridization between similar sequences can make it difficult to interpret results. Consequently, as sequencing has become cheaper RNA-seq technologies (Mortazavi et al., 2008; Wang et al., 2009) are starting to replace DNA microarrays.

1.4.3 Chromatin immunoprecipitation

Beyond the static genome lies a dynamic collection of proteins and modifications that decorate the genome. As described above, in this thesis we will want to know where TFs bind to the genome and what chromatin modifications exist in each genomic region. Chromatin immunoprecipitation (ChIP; Solomon et al., 1988) followed by the application to a microarray (ChIP-chip; Ren et al., 2000; Iyer et al., 2001) or sequencing (ChIP-seq; Robertson et al., 2007) permit the assaying of these genomic features in a dynamic manner (Figure 1-4). The same type of arrays can be used for ChIP-chip as are used for expression analysis. However, because there are many more genomic features that are candidates for protein binding compared to the number of mRNAs, producing arrays for ChIP-chip can be a much more challenging problem and require trade-offs in terms of coverage and cost.

A number of potential issues exist with both ChIP-chip and ChIP-seq. First, they inherit the problems with their underlying technology and can have a significant error rate (Buck and Lieb, 2004). For ChIP-seq, repetitive regions and sequencing errors make mapping the reads to the genome can be a significant challenge (Park, 2009).

Beyond technical issues, ChIP techniques, when applied to TFs, are inherently only able to find regions bound in the specific sample and are unable to find other potential binding sites for a factor. Further, the high rate of turnover between species (Odom et al., 2007) suggests that many binding sites may not be selectively functional making the results of ChIP experiments difficult to put in context. Finally, they are unable to distinguish between sequence specific binding and non-specific binding due to highly accessible regions, which may constitute a large number of the bound regions (Li et al., 2008).

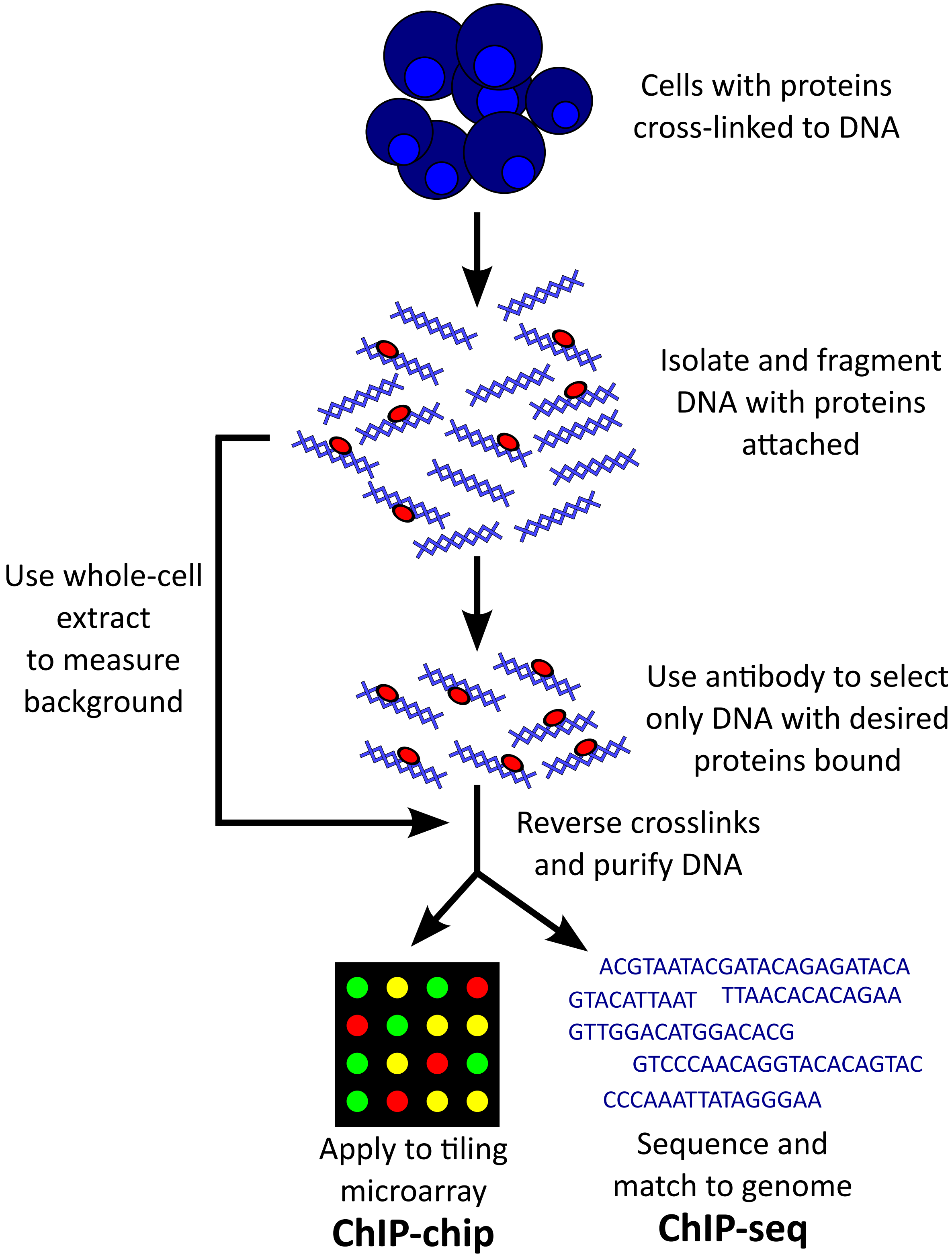

Cells with proteins cross-linked to DNA
Isolate and fragment DNA with proteins attached
Use whole-cell extract to measure background
Use antibody to select only DNA with desired proteins bound
Reverse crosslinks and purify DNA
ACGTAATACGATACAGAGATACA
GTACATTAAT
TTAACACACAGAA
GTTGGACATGGACACG
GTCCCAACAGGTACACAGTAC
CCCAAATTATAGGGAA
Apply to tiling microarray
ChIP-chip
Sequence and match to genome
ChIP-seq

Figure 1-4: Diagram of chromatin immunoprecipitation (ChIP) followed by assessment with microarray (chip) or sequencing (seq).

1.5 Common data used

1.5.1 Comprehensive collection of known motifs

Known motifs were collected primarily from large scale datasets or databases, but with significant manual annotation. For human, we collected human, mouse, and rat motifs from Transfac (version 11.3; Matys et al., 2003), vertebrate motifs from Jaspar (version 2008; Sandelin et al., 2004), and large scale systematic motifs generated by Protein Binding Microarrays (Berger et al., 2006; Badis et al., 2009; Berger et al., 2008). For *Drosophila* we used fly motifs from Transfac and Jaspar in addition to motifs collected from various literature sources (Sen et al., 2010; Reed et al., 2008; Noyes et al., 2008a,b; MacArthur et al., 2009; Down et al., 2007; Ivan et al., 2008; Wasserman and Sandelin, 2004). Names for the motifs were standardized by factor name (in human some families were collapsed if their motifs were similar enough; fly motifs were named by their fly base symbol). Hierarchical clustering of mammalian motifs is performed using centroid linkage and a cutoff of 0.95. This cutoff is high enough where the motifs essentially match the same genomic locations and is used only for identifying redundancy. For each cluster only the motif closest to the centroid is retained.

1.5.2 Genome annotations

Because the work in this thesis is centered around model organisms, we are able to exploit available annotations. When performing motif instance prediction, it is important to exclude regions that may have other sources of evolutionary constraint (e.g., coding sequence) and regions that are difficult to align (e.g., repeats). Consequently, all simple repeats, repeat masked regions, coding regions, $3'$ UTRs, exons from non-coding genes, and chromosomes Y and M are excluded unless otherwise specified. Simple repeats and repeat masked regions are taken from UCSC for the appropriate assembly (Kent et al., 2002). Fly gene annotations for dm3 were taken from Flybase v5.28 (Tweedie et al., 2009) and miRBase v15

(Griffiths-Jones et al., 2008). Human hg18 annotations (used in Chapters 2 and 5) are taken from GENCODE v2b (Harrow et al., 2006); hg19 annotations (used in Chapter 3) are taken from Gencode v4.

1.6 Thesis overview

This thesis deals with using computational approaches to better understand gene regulation. The contributions include:

- A novel, practical algorithm for predicting comparative motif instances (Chapter 2). We analyze the performance of the algorithm in recovering motifs in the context of experimental and functional datasets and for differing numbers of species. We conclude that additional species will allow us to predict additional instances at the same confidence. To make this analysis possible, we develop a number of high-performance computational tools.

- The systematic annotation of motifs for hundreds of human ChIP-seq datasets (Chapter 3), appropriate for use with the method developed in Chapter 2. We use statistical corrections for enrichment and carefully chosen controls to correct for various issues and are able to find: (1) the most accurate motif for a TF; (2) a handful of unvalidated, novel motifs; (3) cooperating and antagonizing factors; and (4) meaningful differences in binding of the same factor between cell types. We do a thorough analysis of the results and find many factor relationships that are confirmed in the literature and make several additional predictions appropriate for follow-up.

- Methods for computationally predicting microRNA (miRNA) hairpins and their corresponding $5'$ cleavage sites using comparative and structural information (Chapter 4). To predict these regulators we use a customized random forest algorithm and achieve over 4,500-fold enrichment for real hairpins over random hairpins in the genome. We find that our performance is better than a competing algorithm that was run on the same data. Predicting the mature miRNA produces additional motifs for use with our motif instance algorithm; we use an SVM and update several previously made predictions, leading to a significant update in the target spectra.

- The annotation of cell line specific factors in human and fly using chromatin modifications (Chapter 5). We find: (1) our comparative motif instances can be reliably used to predict key regulators of cell types; (2) these regulators can be classified as activators or repressors by how their enrichment signatures correlated with the expression of the regulators in the same cell types; and (3) the enrichments of activator motifs and their correlation with expression can be used to classify chromatin marks or states in terms of activator potential.

- The systematic testing of the predictions made in Chapters 2 and 5 (Chapter 6). We apply a massively parallel reporter assay (MPRA) to measure the enhancer activity of thousands of sequences centered on motif instances and their engineered manipulations. In doing so, we significantly increase the number of experimentally validated enhancers and careful statistical analysis leads to a number of insights: (1) 145-bp is often sufficient to capture the enhancer activity when centered on motif instances; (2) enhancers centered on comparative motif instances are $\sim$2 times more likely to be functional as those centered on random motif matches; (3) several other properties are correlated with sequences that have strong enhancer activity, including chromatin mark dip scores (an indication of nucleosome exclusion), motif match strength, and the enrichment of motifs for other factors; (4) manipulating the motif match affects expression consistent with the specificity indicated by the PWM: disruptive mutations that would prevent TF binding eliminate enhancer activity, whereas mutations permitted by the PWM do not affect enhancer activity; (5) disrupting the binding sites for repressors can lead to an increase in expression in the cell type where the repressor in active; and

(6) together these results validate our motif instances and our factor/cell-line predictions.

Chapter 2

Regulatory motif instance prediction

This chapter will describe and evaluate an algorithm for predicting functional motif instances using multiple, closely related species. We define functional motif instances as those that would result in a reduction of fitness of an organism if disrupted, although we also expect them to be more likely to be biochemically active than a simple motif match. The comparative motif instances produced here will be used in Chapter 5 to examine the relationship between motif instances and chromatin modifications and then systematically experimentally validated in Chapter 6. While I was responsible for almost all aspects of the implementation and analysis, some of this work, particularly the initial algorithmic design and the *Drosophila* results, were done as part of a collaboration with Alexander Stark. This chapter is based on results previously published in Kheradpour et al. (2007), Stark et al. (2007b), and Lindblad-Toh et al. (2011) with notable additions.

2.1 Introduction

Once the motif for a regulator has been determined, a natural desire is to predict its functional locations. However, a consequence of short nature of most metazoan motifs (5-15 bp) is that they will frequently match the genome just by chance — a fully specified 6-mer will match a uniformly random genome once every $4^6 \approx 4000$ bases. A large mammalian genome therefore contains hundreds

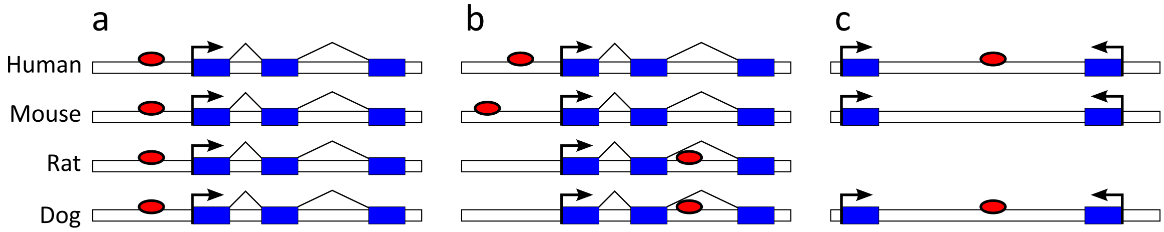


Figure 2-1: Challenges associated with motif instance identification using many aligned genomes (hypothetical motif matches are indicated in red). (a) The simple, straightforward case is when an instance is found fully conserved in the orthologous position near a given gene. (b) Motif turn-around or alignment errors can lead to a motif match being found in a location proximal to one in the target genome, but not directly aligned. (c) Motif matches can be missing due to turn-over or sequencing errors. The motif instances can also be found far from a gene, making them difficult to assign.

of thousands of matches for such a motif, far more than the number of regions bound in an experimental assay (tens of thousands at most). The source of this discrepancy is not completely clear, but chromatin structure, lack of necessary cofactors or motif multiplicity, and incorrect models of binding have been proposed as possible explanations (Wasserman and Sandelin, 2004; Badis et al., 2009).

Consequently, the general approach toward motif instance prediction has been to increase power by looking for motif matches that are more likely to be associated with functionality but less likely to occur just by chance. A popular way to do this is by finding regions of the genome that are enriched for a set of transcription factors known to act in concert (Berman et al., 2002; Schroeder et al., 2004; Philippakis et al., 2006). This has been successful because it requires only one genome but can predict sequences that have a high probability of functionality. However, these approaches are inherently require a set of motifs known to act together and are unable to find motif matches that occur in isolation.

An alternative approach that is able to find isolated binding sites is phylogenetic footprinting, which exploits the preferential conservation of motif instances. Early work in this area mainly focused predicting motif instances that were perfectly conserved in orthologous regions between two or more species (Sharan et al., 2003; Ettwiller et al., 2005; Lewis et al., 2005) whereas Ho Sui et al. (2005)

matched motifs to areas with conservation above some threshold. Conversely, Blanchette and Tompa (2002) used an alignment-free approach to find k-mers in orthologous promoters that were unusually well conserved and Moses et al. (2004) models binding using a strict phylogenetic model to find regions that evolve according to the motif and not the background. These methods were generally not designed to cope with large phylogenies of species containing sequencing and alignment errors.

In this chapter we present our own practical alignment-free phylogenetic footprinting algorithm. We will then evaluate our method separately using 29 placental mammals and 12 fruit flies. We expect this method to be generally applicable as long as whole genome alignments can be produced and there is sufficient total branch length.

2.2 Producing robust comparative motif instances

The complexity of large phylogenies leads to a number of issues that prevent simple matching of motifs to conserved genomic regions. Sequence properties, such as dinucleotide biases, must be considered because they greatly influence the abundance of a motif and its observed mutation rate. Consequently, we produce control motifs specific to each motif we scan that are diverse and have similar properties to our original motif (Figure 2-2). Further, the low coverage genomes used for some studies (e.g., the 29 mammals) will lead to large gaps in the assemblies for some species necessitating a scoring scheme that does not strongly penalize for a missing species in a dense species tree. Even with complete data, unannotated functional elements may match a motif and produce an apparently conserved instance, requiring the measuring of background level of conservation. Because alignment algorithms are imperfect and motif turnover may lead to motifs appearing to have moved (Odom et al., 2007), we support shifts in the placement of motif instances in the alignment. These motif turn-over events represent conservation of function, not a phylogenetic relationship between the

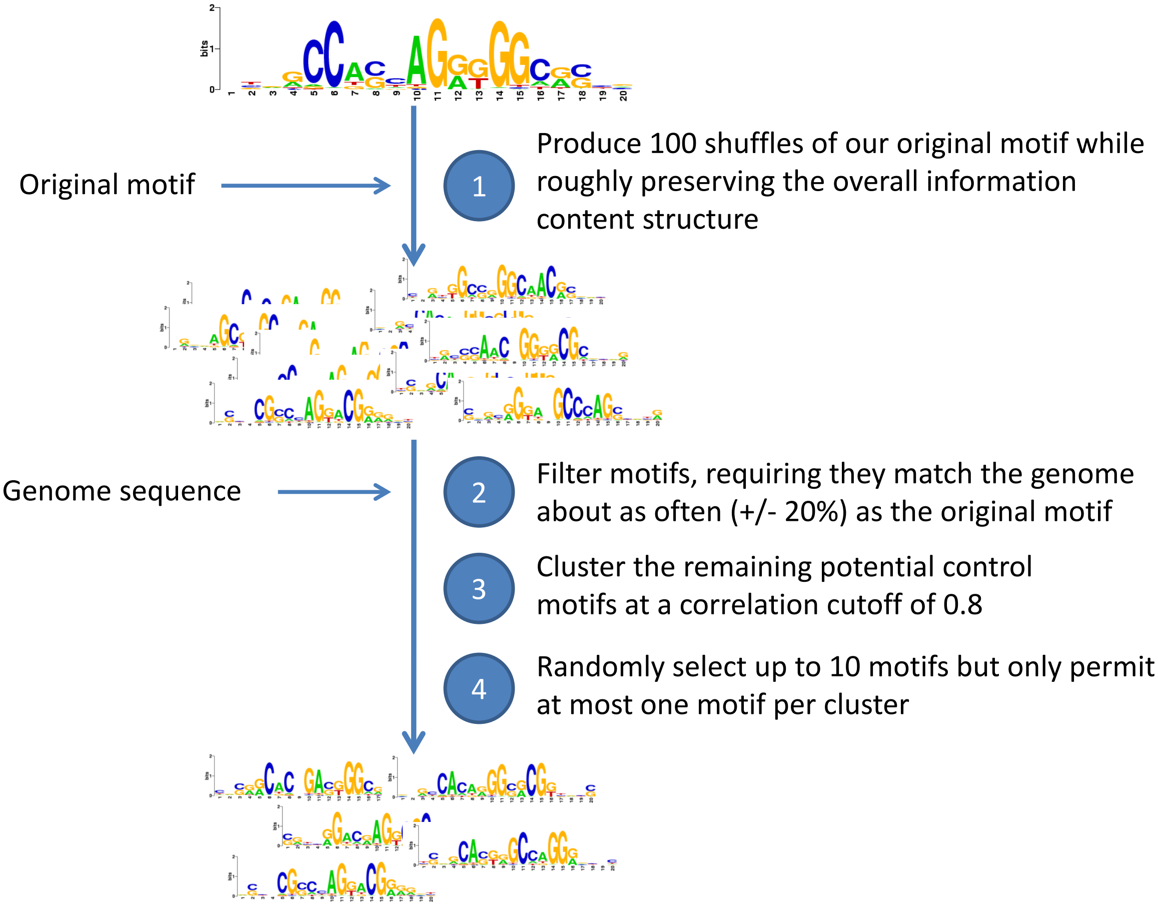

Original motif
1
Produce 100 shuffles of our original motif while roughly preserving the overall information content structure
Genome sequence
2
Filter motifs, requiring they match the genome about as often (+/- 20%) as the original motif
3
Cluster the remaining potential control motifs at a correlation cutoff of 0.8
4
Randomly select up to 10 motifs but only permit at most one motif per cluster

Figure 2-2: Procedure for generating shuffled motifs.

corresponding bases. Consequently, our motif instances are produced using an alignment-free approach that has been used by others for similar purposes (Ward and Bussemaker, 2008).

For each motif in our database (see Section 1.5.1), 100 putative control motifs are generated by randomly shuffling the columns of each PFM (Figure 2-2). Because the particular way the information content of a motif is ordered may affect the background level of conservation (e.g., a group of specified bases surrounded by unspecified bases may be more likely to be conserved by chance), we create three bins of information content and shuffle only within each bin. Each of the 100 shuffled motifs is then matched to the genome (as described in Section 2.3) and only those that have ±20% the number of matches of the original motif are considered (building on Lewis et al., 2003). The remaining motifs are then clustered

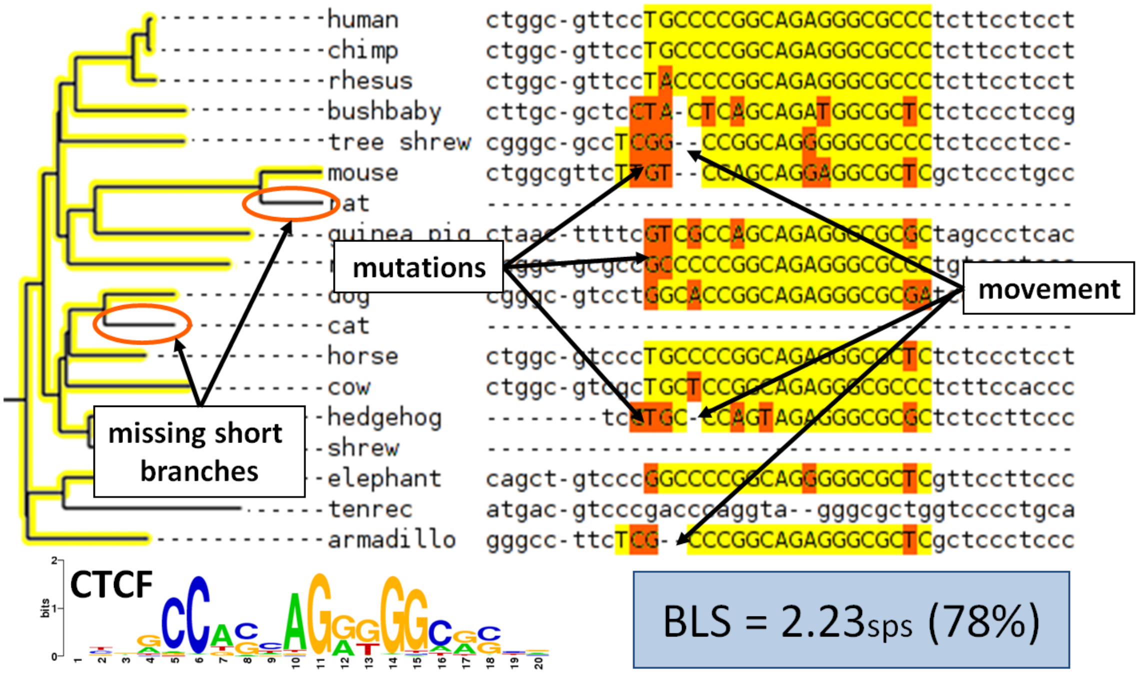

human
chimp
rhesus
bushbaby
tree shrew
mouse
rat
guinea pig
dog
cat
horse
cow
hedgehog
shrew
elephant
tenrec
armadillo
ctggc-gttccTGCCCCGGCAGAGGGCGCCCtcttcctcct
ctggc-gttccTGCCCCGGCAGAGGGCGCCCtcttcctcct
ctggc-gttccTACCCCGGCAGAGGGCGCCCtcttcctcct
cttgc-gctcCTA-CTCAGCAGATGGCGCTCtctccctccg
cgggc-gccTCGG--CCGGCAGGGGGCGCCCtctccctcc-
ctggcgttcTTGT--CCAGCAGGAGGCGCTCgctccctgcc
ctaac-ttttcGTCGCCAGCAGAGGGCGCGCtagccctcac
cgggc-gtcctGGCACCGGCAGAGGGCGCGA
ctggc-gtcccTGCCCCGGCAGAGGGCGCTCtctccctcct
ctggc-gtcgcTGCTCCGGCAGAGGGCGCCCtcttccaccc
--------tcaTGC-CCAGTAGAGGGCGCGCtctccttccc
cagct-gtcccGGCCCCGGCAGGGGGCGCTCgttccttccc
atgac-gtcccgacccaggta--gggcgctggtcccctgca
gggcc-ttcTCG-CCCGGCAGAGGGCGCTCgctccctccc
mutations
movement
missing short branches
CTCF
bits
BLS = 2.23sps (78%)

Figure 2-3: Example motif match to CTCF and corresponding computation of BLS. BLS is equal to the size of the smallest subtree that contains all the species with a motif match. The advantages of this approach over a simple measure of conservation are indicated. This example is illustrative and not all species used for 29 mammals study are included. BLS is measured here in substitutions per site (sps).

at a 0.8 correlation cutoff and up to 10 control motifs are chosen in random order, allowing only one motif per cluster. We find that the cutoff of 0.8 results in motifs that are sufficiently dissimilar as to not frequently match the same sequences disrupting our statistics which assume independence.

Because we require identical base-composition and similar number of overall matches, for some motifs no control motifs can be generated and thus are not amenable to our algorithm (this generally occurs for $<$ 1% of motifs). The algorithm does not permit lower quality controls and thus does not compromise the quality of motif instances.

Our analysis is centered around the same species as an input whole genome alignment, which is typically a model organism such as human, mouse, or *Drosophila melanogaster*. All the other species in the alignment are used as informants. For each motif match (Section 2.3) in the target genome, we compute a branch length score (BLS; Figure 2-3)). This is done by using whole-genome alignments to identify the other species that have a motif match in the aligned position (expanding this to allow motif movement is described below). The BLS is then defined as the branch length of smallest subtree containing all species with a motif match.

We then produce a mapping between BLS and confidence (intended to approximate 1 - false discovery rate; Figure 2-4) for each BLS (at 100 evenly spaced values from 0 to the total branch length of the tree) by computing the number of instances that reach that BLS score and comparing that to the number we would expect to be according to the control motifs. Let $\bar{r}_b$ be the fraction of instances for control motifs that have BLS score $\geq b$, and let $r_b = \frac{n_b}{n_0}$ where n_b indicates the number of motif matches to our motif that have BLS $\geq b$. We define the confidence c_b:

$$\begin{aligned} c_b &= \frac{n_b - \bar{r}_b \times n_0}{n_b} \\ &= 1 - \frac{\bar{r}_b}{r_b} \end{aligned}$$

notice that while c_b will be negative if the control motifs are more conserved than the original motif, because $c_0 = 0$ and will always have the most instances, we never report motif instances with less than 0 confidence.

This measure of alignment-free conservation does not use a fully specified model of evolution. However, because it empirically corrects for phenomena that would otherwise be difficult to model (e.g., conservation due to non-coding RNAs), we have found it to be useful in practice, which we will show in the remainder of this thesis.

Wilson score interval (Wilson, 1927) with $z = 1$ is applied to both r (correcting downward) and r_c (correcting upward) in order to produce a conservative estimate of confidence in situations with few instances. This is essentially the same computation we use for evaluating enrichments (see Section 5.2). We also permit motif movement by repeating the procedure for each of the 32 windows

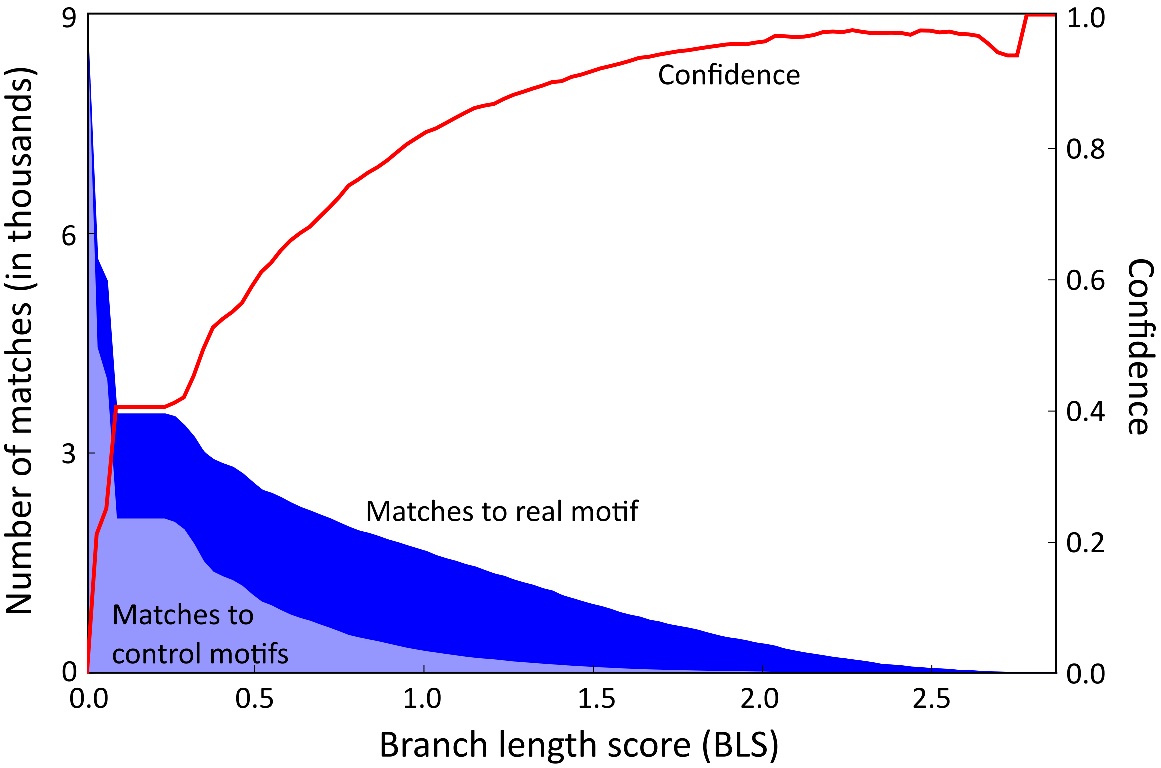


Figure 2-4: Computation of confidence score for motif instances of CTCF in mammals. The number of instances for the motif that reach each branch length is computed (dark blue). Control motifs are used to compute an expected background level (light blue), correcting for alignment-free conservation by chance or due to overlap with unannotated elements. The fraction of the dark to light blue above results in the confidence score (red).

$w = 0, 5, 10, 20, \ldots, 100, 120, \ldots, 500$ allowing both the motif and the control motifs to move w bases in the informant genomes relative to the position aligned to our target species. Consequently, for each confidence cutoff from $0.1, 0.2, \ldots, 0.9$ the BLS and w combination that results in the highest sensitivity is chosen. However, permitting movement only modestly increases the number of instances (∼17% in human) and consequently our method is still largely alignment driven (Figure 2-5).

In human, confidence prediction is done on only autosomes (non-X/Y/M chromosomes), and then instances are produced on the chromosome X using the mapping produced on the autosomes but with a tree produced on chromosome X. This is important to correct for the higher background level of conservation of chromosome X (Vicoso and Charlesworth, 2006). Chromosome Y is ignored because data in other species is incomplete (not all sequenced mammals were male) . Scaling and motif movement analysis shown below ignores instances on chromosome X (Figures 2-5, 2-10, 2-14 and 2-15).

2.3 Techniques for matching motifs

Typically, motifs are available as 4xN matrices indicating for each position the frequency of each base (position frequency matrix; PFM). Before matching, PFMs are typically converted to PWMs by incorporating a background model and putting them into log space. In this thesis we use a pseudo count of 0.001 (to prevent undefined values) and a uniform background. In this chapter, we will generally use a threshold corresponding to 4^{-8} as determined by TFM-Pvalue (Touzet and Varre, 2007). Matching a single PWM to one genome is a straightforward computational task involving summing floats across each position and comparing to a cutoff. A number of tools are available for this task, including MAST (Bailey and Gribskov, 1998), storm (Schones et al., 2005), and AffinityProfile (Foat et al., 2006). For our purposes we needed a tool that could: (1) match motifs to multiple species, (2) scan in a window in the other species while avoiding the same match

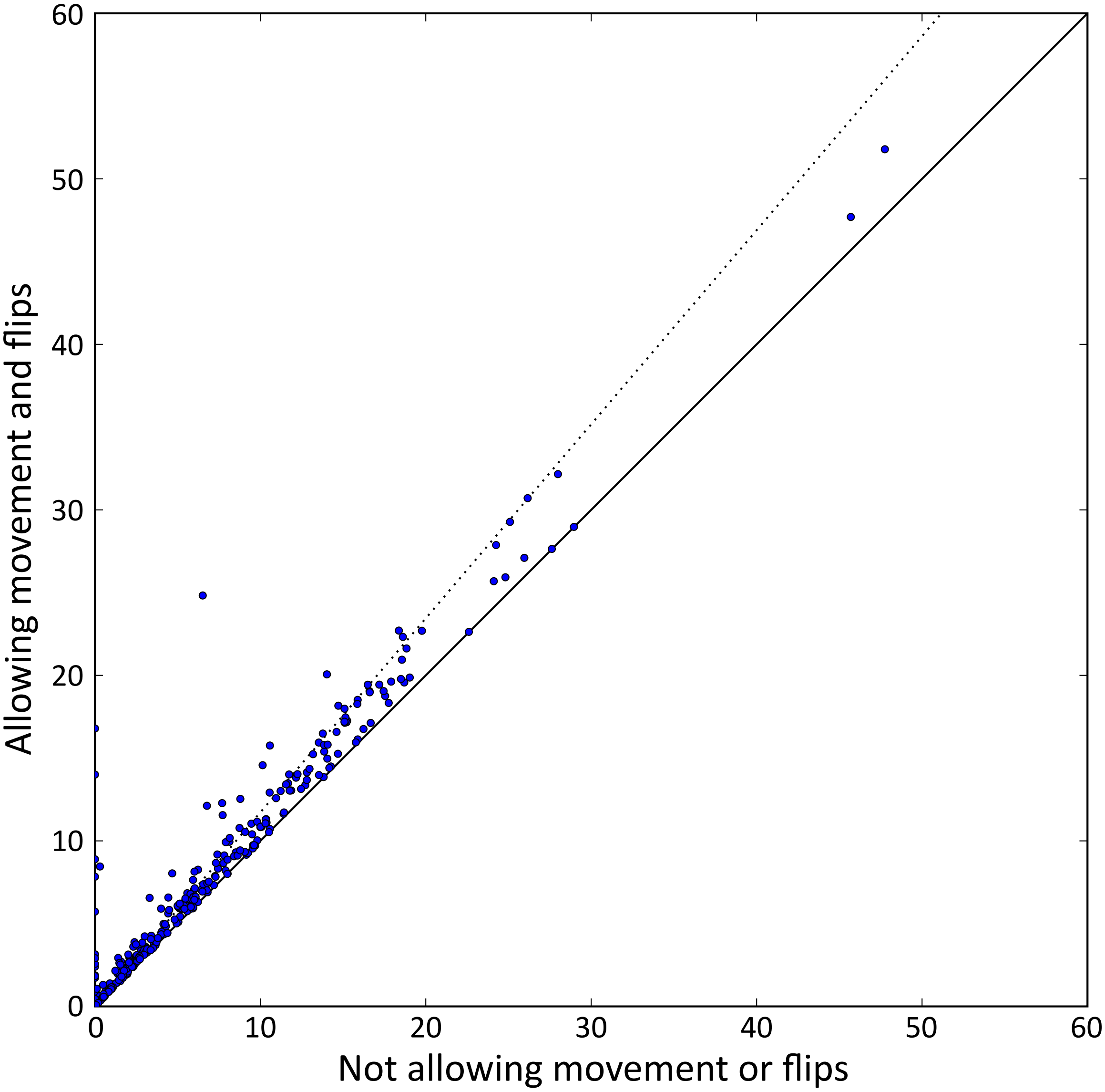


Figure 2-5: Number of motif instances (in thousands) for each human motif at 40% confidence when allow motif movement and flips or not. Solid line indicates no change whereas dotted line is a 17% increase (the overall proportion of additional instances).

being assigned to multiple matches in the target, while being (3) fast enough to feasibly match thousands of motifs to mammalian scale genomes. Here we describe technical details of our motif matching software (written in C) that fulfills these requirements.

For each motif, all fully specified 8-mers that could begin a motif match are computed (recursively; motifs less than length 8 are padded with Ns). These are put into a 4^8 entry lookup table so that while scanning the target species genome a single lookup can produce a list of all potentially matching motifs. We found that this heuristic dramatically increases the speed of matching (about 10 times, for our typical mammalian runs).

Once we have found a motif match in the target species, we must determine which informant genomes also match. For this purpose, we compute the flanking matches to our motif in the target genome. These are used to eliminate the aligned regions in the informant species that are closer to some other motif match in the target species. This is an important step to avoid a single match in an informant species from making multiple target species matches appear conserved. Once the informant species with motif matches are determined, we compute the BLS using a parent tree representation of the phylogenetic tree.

Finally, many analyzes require computing enrichments by counting the number of instances in each type of region. Because the resulting match files can be as large as 200 gigabytes compressed, most software to produce overlaps would fail when trying to load them into memory. To deal with this challenge we produced software to determine overlapping regions for files sorted by chromosome then start position. Overlaps are then produced using the following algorithm (which is the same as the "chromsweep" algorithm independently implemented by BEDtools; Quinlan and Hall, 2010):

```
stored_lines = {}
while new_line = read_line(file_1):
    // when comparing positions, also compare chromosomes

    // scan through list of stored elements and remove ones
    // we will never overlap in another read in file_1 line
    for line in stored_lines in order:
        if line.end < new_line.start:
            delete stored_lines[line]
        else if line.start >= new_line.start:
            break // short circuit

    // read in new lines
    while not(eof(file_2)) and
        (
            isempty(stored_lines) or
            stored_lines.last.start <= new_line.end
        ):
        line = read_line(file_2)
        if line.end >= new_line.start:
            stored_lines.push(line)

    // print out matching elements
    for line in stored_lines in order:
        if line.start <= new_line.end:
            print new_line, line
        else:
            break // short circuit
```

The running time of this algorithm is $O(f_1 + f_2 + o)$ where f_1 and f_2 are the number of lines in each file and o is the number of overlaps (we only examine a constant number of stored_lines more than we print or read in each iteration). More importantly, memory usage is only $O(o_{\text{max}})$ where o_{max} is the maximum number of elements a region in f_1 overlaps in f_2. Sorted motif matches are automatically produced by scanning chromosomes in alphanumeric sort order and sorted regions are produced by unix sort (which, depending on implementation,

Confidence	No. motifs reaching confidence	Total No. instances	% examined bases covered	No. TFs with a motif reaching confidence	Total No. instances (best motif per TF)
0.0	630	55,021,406	80.6	335	35,366,716
0.1	540	15,817,545	45.3	294	11,181,918
0.2	492	8,385,913	26.0	270	6,068,955
0.3	435	4,697,272	14.3	252	3,495,271
0.4	375	2,675,802	7.7	225	2,050,302
0.5	293	1,449,752	3.9	188	1,175,237
0.6	216	707,141	1.7	151	595,984
0.7	129	269,944	0.6	101	240,849
0.8	56	90,464	0.2	45	80,138
0.9	16	33,822	0.1	14	29,080

Table 2-1: Basic statistics on the predicted human motif instances.

uses disk cache as necessary). Together these tools permit us to match hundreds of motifs to entire mammalian genomes with relative ease (on an appropriately sized cluster).

2.4 Validating the predicted mammalian regulatory network

Of the 688 motifs initially in our database, 630 (representing 335 factors) were able to be matched at the required stringency and have at least one shuffle motif. The number of instances found at various cutoffs are indicated in Table 2-1.

ChIP-chip/seq is a popular experimental technique for determining the binding sites of a factor *in vivo* (Section 1.4.3). However, because ChIP inherently does not capture binding events that occur in all cell types, and not all binding events are conserved, we do not expect perfect concordance between ChIP regions and comparative motif instances. Regardless, an enrichment in one relative to the other would suggest the validity of the motif instances because we expect functional motif instances to be more likely to be bound *in vivo*.

For this purpose, we assembled a database of published ChIP-chip/seq

Factor	Cell type	Technology	Num peaks	Motif used	Citation
CTCF	CD4+ T ES (mouse)	Sequencing	21,544 8,546 (+mouse)	Jaspar MA0139.1	Barski et al. (2007) Mouse: Goren et al. (2010)
ER	MCF-7	Paired-end Tags	1,229 7	Transfac M00191	Lin et al. (2007a)
Fos	K562 CML	Sequencing	18,963 7	Transfac M00926	Raha et al. (2010)
FOXA2	Liver	Promoter array	143 19 (+mouse)	Jaspar MA0047.2	Odom et al. (2007)
HNF1	Liver	Promoter array	246 23 (+mouse)	Jaspar MA0046.1	Odom et al. (2007)
HNF4	Liver	Promoter array	1,231 99 (+mouse)	Transfac M01036	Odom et al. (2007)
HNF6	Liver	Promoter array	149 20 (+mouse)	Transfac M00639	Odom et al. (2007)
Myc	K562 ES (mouse)	Sequencing Promoter array (mouse)	15,749 2,399 (+mouse)	Transfac M00187	Raha et al. (2010) Mouse: Kim et al. (2008)
NF-κB	GM12878	Sequencing	38,559	Jaspar MA0061.1	Kasowski et al. (2010)
NRSF	Jurkat T	Sequencing	1,931	Transfac M00325	Johnson et al. (2007)
p53	HCT116	Paired-end Tags	62,939	Transfac M00034	Wei et al. (2006)
STAT1	HeLa-S3	Sequencing	41,530	Transfac M00224	Robertson et al. (2007)
YY1	NT2/D1	Sequencing	11,018	Transfac M00651	Consortium (2011a)

Table 2-2: Listing of datasets and motifs used in human analysis. Datasets were identified from the literature and the peaks identified in the study were used after mapping to the appropriate assembly (if necessary) using liftOver (Kent et al., 2002). For factors that also had a dataset available in mouse, we show the number of peaks found in human that were bound in the orthologous mouse positions. When multiple motifs were available for a factor, we chose the one with the highest enrichment in the human dataset (ignoring conservation).

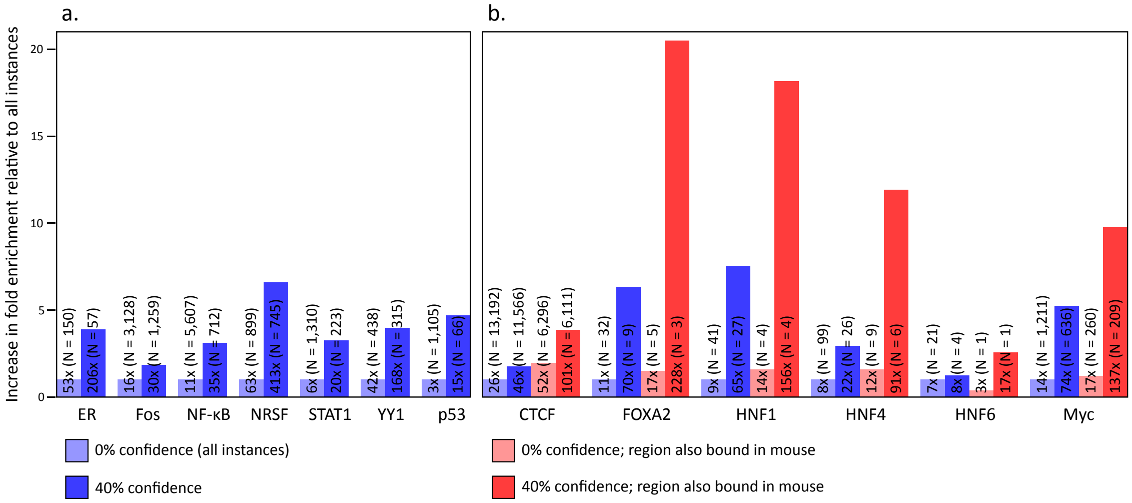


Figure 2-6: Enrichment of motifs in published experimental datasets. (a,b) Known motifs for each factor show an enrichment in experimental datasets which increases with alignment-free conservation. (b) Enrichment dramatically increases for regions that are bound in both human and in the orthologous positions in mouse.

datasets (Table 2-2) and we do, indeed, observe increased enrichment of our comparative motif instances (Figure 2-6). Enrichments are computed as the ratio of the fraction of motif instances inside a region to the fraction of bases inside that region. For example, if 20% of a motif's instances are bound by a given factor, but only 1% of the genome is bound, then we would report an enrichment of 20-fold. Moreover, this enrichment increases, often substantially, with increasing confidence (Figure 2-7) and is also seen for fly datasets (Figure 2-8), demonstrating the generality of this method.

In many cases, a larger proportion of the predicted motif instances do not overlap experimentally bound sites than is expected given the precision indicated by the confidence level. These may result from (1) an inaccurate confidence prediction, (2) false negatives in the experimental procedure, (3) the existence of a regulator with similar binding affinity, (4) regions bound in cell types not assayed, or a combination therein. It is, therefore, interesting to note that for CTCF, which has fairly consistent binding across cell types (Kim et al., 2007; Cuddapah et al., 2009) and whose binding specificity is not similar to that of any other factor

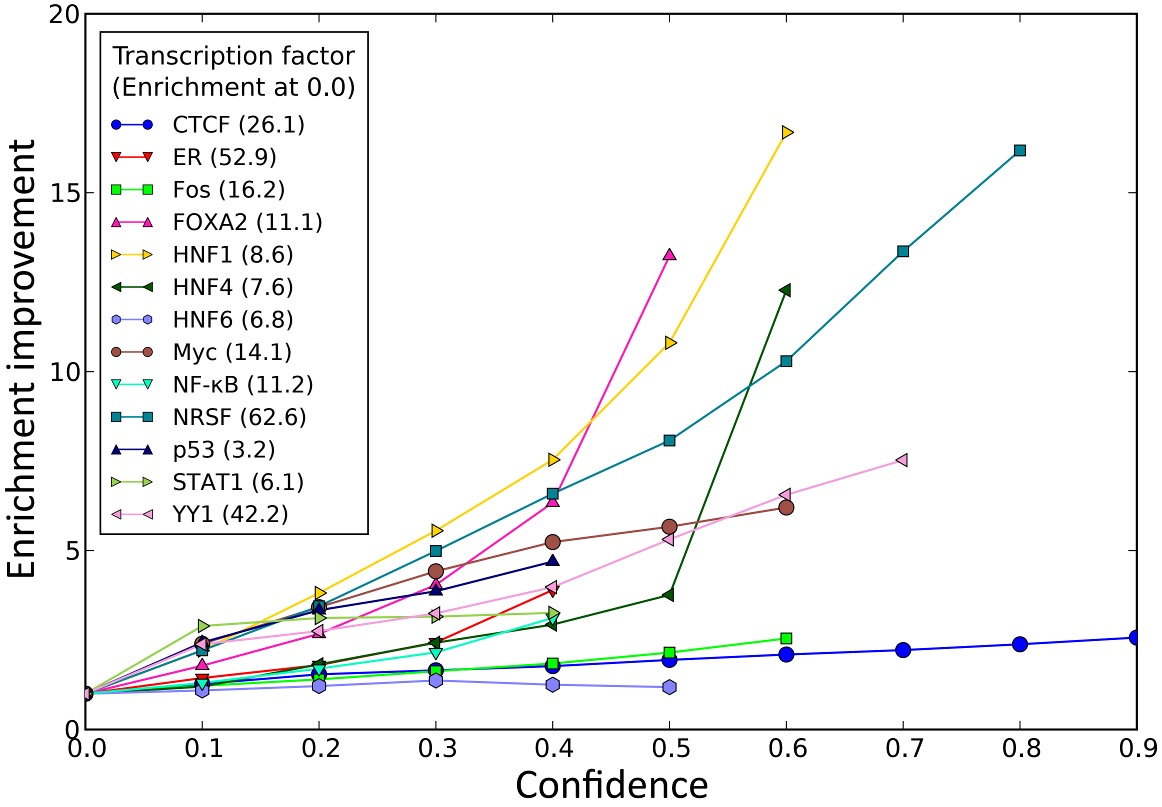


Figure 2-7: Enrichment of corresponding motifs in bound regions. Most factors show consistent and substantial increases in enrichment with increasing confidence. Motif enrichments divided by enrichment at 0.0 confidence (i.e. all motif instances).

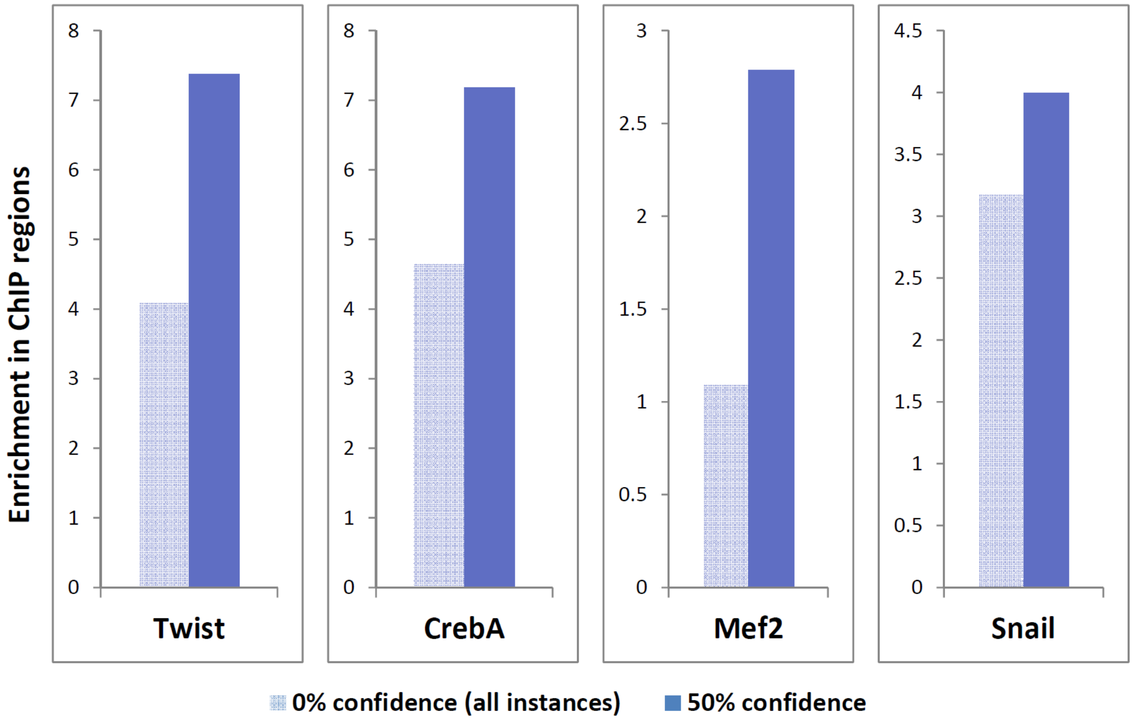


Figure 2-8: Increase in enrichment is also seen for fly factors with experimental data taken from literature (Abrams and Andrew, 2005; Zeitlinger et al., 2007; Sandmann et al., 2006, 2007).

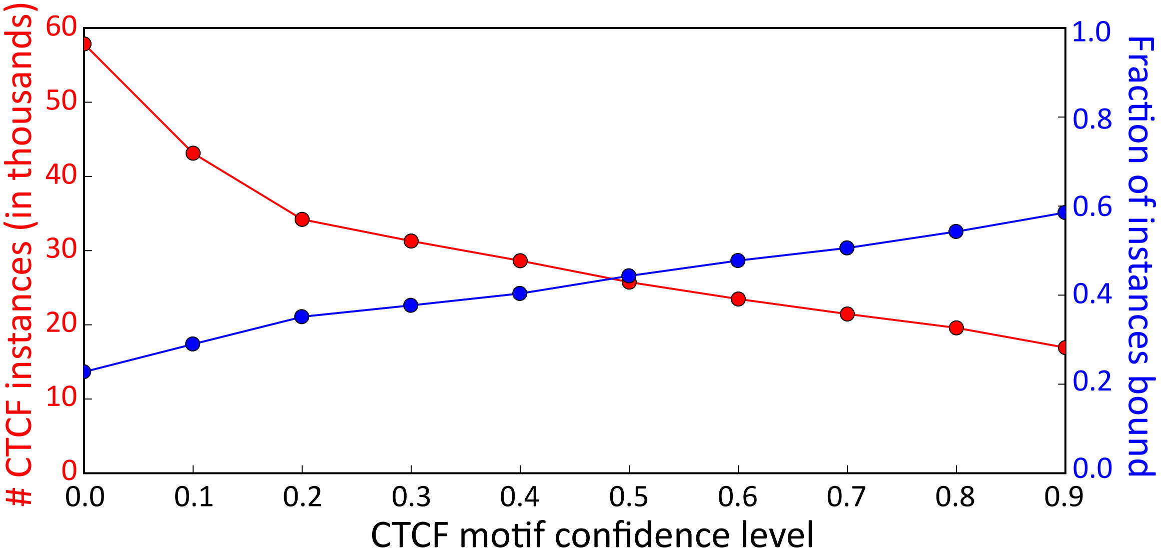


Figure 2-9: The CTCF motif in human has confidence levels roughly tracking the fraction of bound instances (blue; right) while maintaining tens of thousands of instances (red; left).

(maximum motif similarity to any factor lower than that of 92% of factors), has relatively strong agreement between the confidence score and the observed fraction of instances overlapping experimentally identified sites while maintaining a high enrichment (Figure 2-9). A similar trend is seen for NRSF.

There is a substantial difference in the enrichment levels seen for each factor, both for comparative instances and all motif matches. This may be due to a variety of reasons, including: (1) the quality of each experimental dataset and the corresponding known motif; (2) the specificity the factor has for its own motif, versus other contributors of binding; and (3) a variable range of motif turnover depending on the selective pressures on binding sites for a given factor. Despite this, we do consistency see significant enrichment of the motif matches which then increases as we apply conservation.

Because not all binding events are conserved across species (Borneman et al., 2007; Schmidt et al., 2010) and many are not functional (MacArthur et al., 2009), not all experimentally identified regions are expected to have a conserved motif instance. However, the conserved binding sites appear to be very important and indeed tend to be found near targets known to be developmentally important

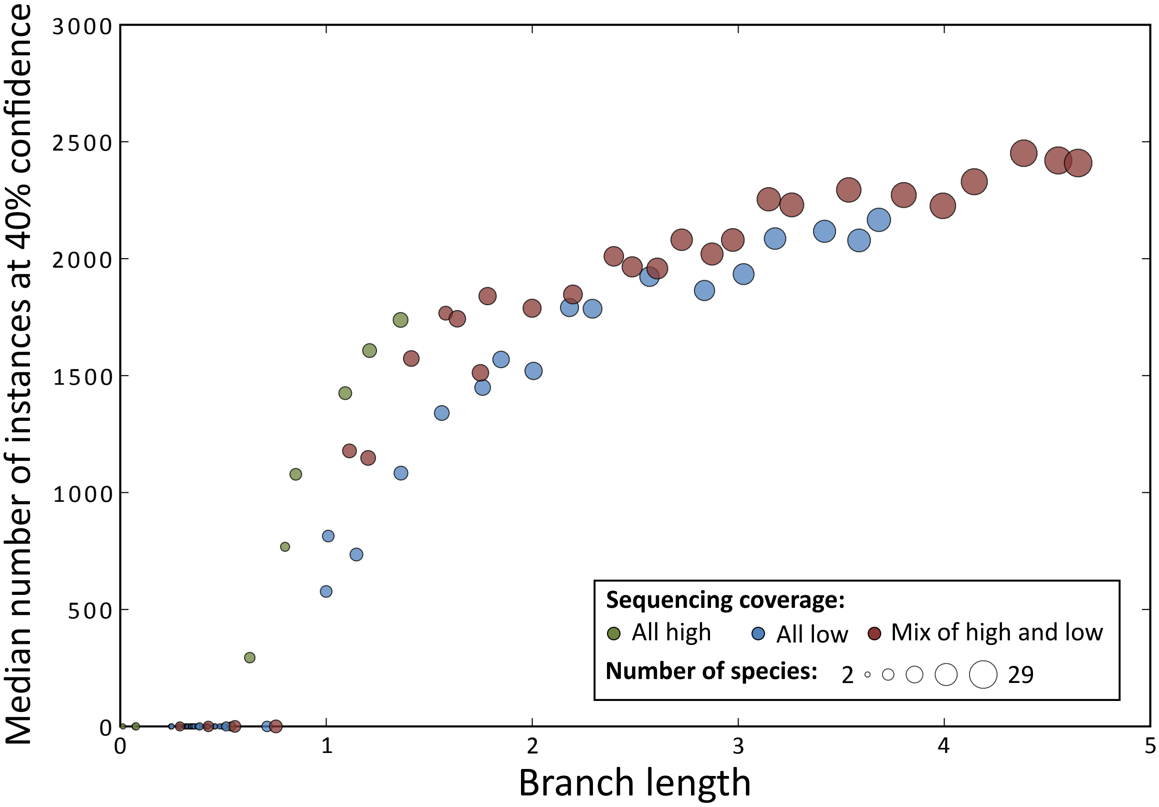


Figure 2-10: Scaling of motif instances using different species subsets. We reran our instance prediction procedure using varying number of species and found that it appears prediction power has not yet saturated. We note that we continued to use the same alignments, simply using only rows corresponding to the species of interest. Consequently, performance strictly on fewer species could be worse because the alignments would not benefit from the intermediate species. The relative value of low coverage (∼2x) and high coverage (∼8x) is shown.

(Schmidt et al., 2010). When we consider only those binding sites that are also conserved in mouse, we find that the enrichment is dramatically higher (Figure 2-6b).

2.5 Scaling of motif instance prediction

Due to their short length and degeneracy, instance prediction greatly benefits from additional species. Indeed, for the mammals we find an increase in the number of predicted instances as the total branch length of the species considered is increased (Figure 2-10). Further, while for a given branch length using high coverage genomes consistently leads to higher sensitivity than using low coverage genomes adding low coverage genomes still significantly improves performance.

We also examined the extent to which adding additional species affected the quality of the motif instances in terms of predicting bound regions. We found that while we found many more instances at the same confidence level using more species, the enrichments in experimental datasets is comparable (Figure 2-11). This is consistent with our expectation that additional species gives us higher power for the prediction of motif instances that are likely to be functional, while our confidence measure accurately assesses their quality.

2.6 ChIP vs. motif instances

While there is a correlation between ChIP and comparative motif instances, there are also significant differences. This raises the question of which more accurately predicts motif instances that are likely to be functional. Consequently, we identified regions independent from both ChIP and motif matches that we expect to be associated with likely functional regions bound by a given factor and use these to compare the two.

We expect instances of NF-κB, an important immune regulator, to be preferentially located in GM/HUVEC enhancers and in the upstream regions of immune response genes. Indeed, we see this trend for both NF-κB motif instances and ChIP bound regions (Figure 2-12). Further, we continue to have an enrichment (4.2-fold) for motif instances that do not intersect with the ChIP regions. Moreover, considerably higher enrichment is seen for motifs in the promoters of immune response genes. ChIP regions that have a motif instance are more than two-fold more enriched in these likely functional regions than either of the criteria alone, demonstrating the complementarity of motifs and experimental techniques. We see a similar trends for STAT1 and p53 in enhancers where we expect them to be active (Figure 2-12).

In fly we do a similar analysis and find that two mesodermal activators (Twist

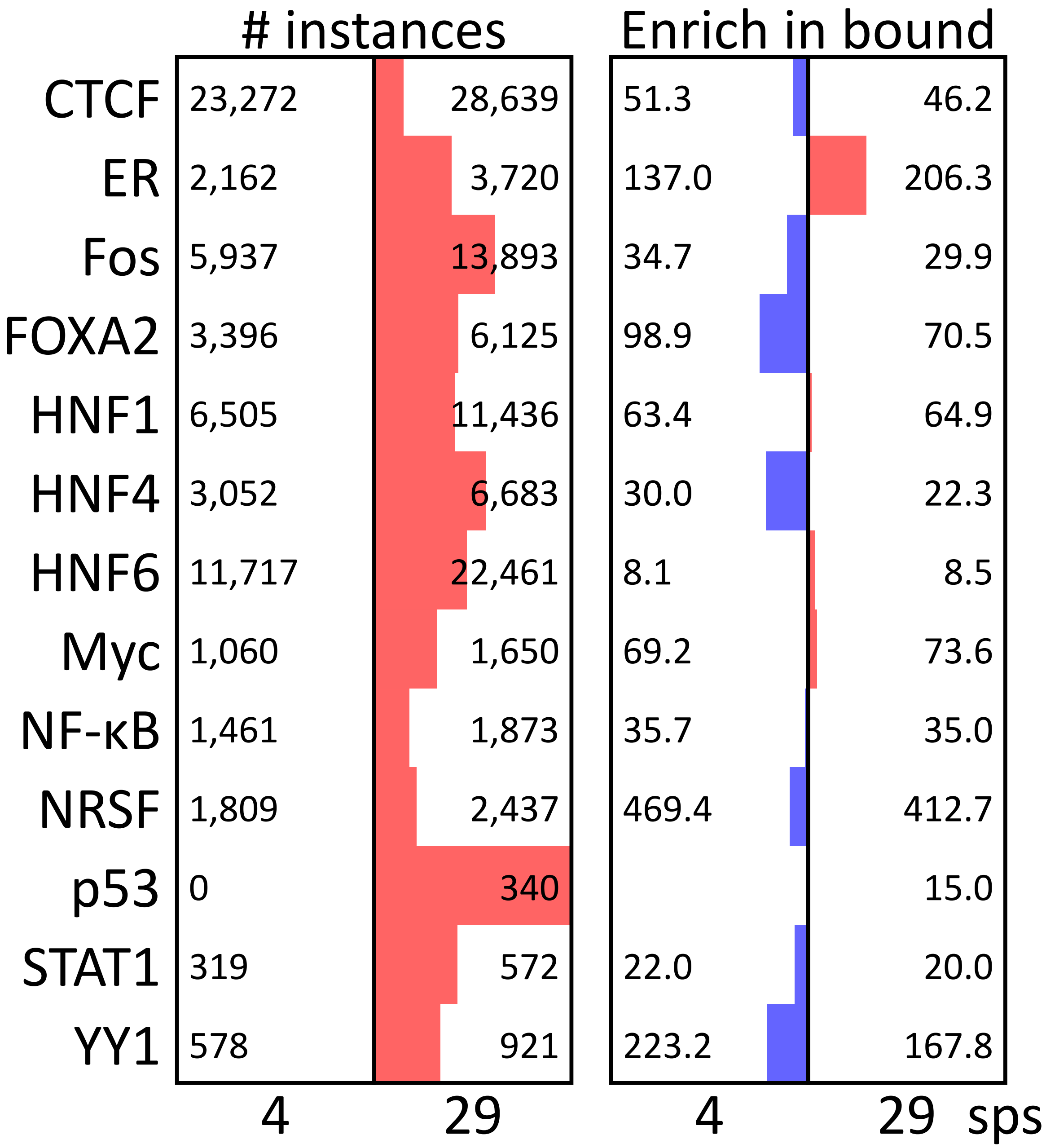
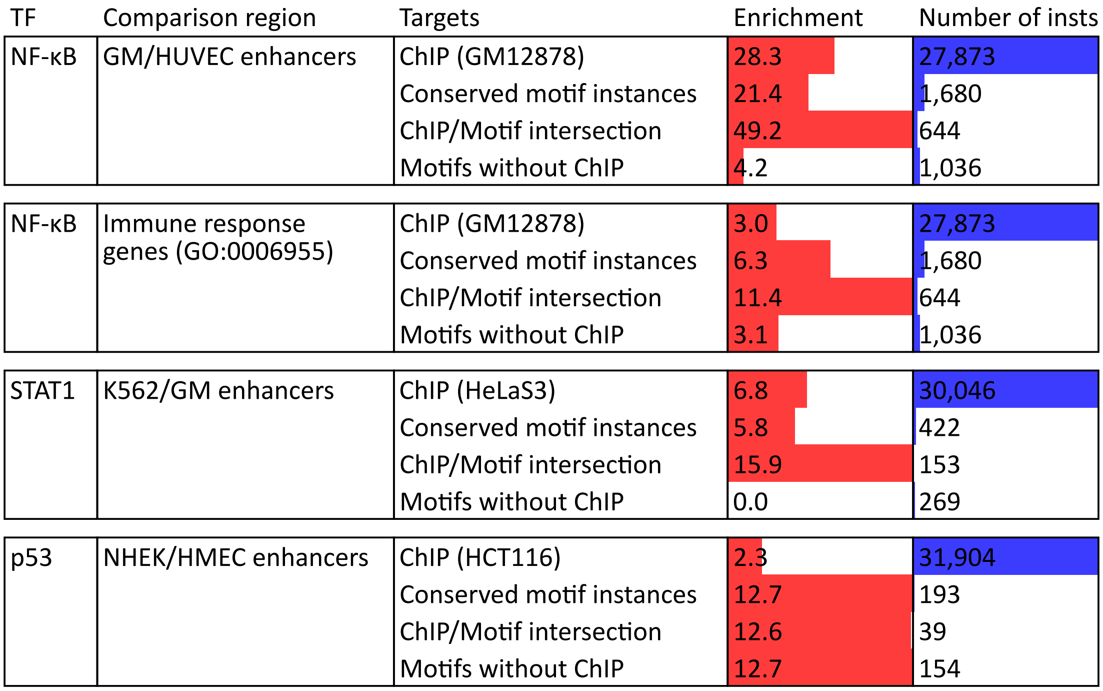


Figure 2-11: Comparison of motif instances at 40% confidence using only 4 mammals (human, mouse, rat, dog) to those found using all 29 mammals. Bars show log ratio of indicated numbers.



TF	Comparison region	Targets	Enrichment	Number of insts
NF-κB	GM/HUVEC enhancers	ChIP (GM12878)	28.3	27,873
		Conserved motif instances	21.4	1,680
		ChIP/Motif intersection	49.2	644
		Motifs without ChIP	4.2	1,036
NF-κB	Immune response genes (GO:0006955)	ChIP (GM12878)	3.0	27,873
		Conserved motif instances	6.3	1,680
		ChIP/Motif intersection	11.4	644
		Motifs without ChIP	3.1	1,036
STAT1	K562/GM enhancers	ChIP (HeLaS3)	6.8	30,046
		Conserved motif instances	5.8	422
		ChIP/Motif intersection	15.9	153
		Motifs without ChIP	0.0	269
p53	NHEK/HMEC enhancers	ChIP (HCT116)	2.3	31,904
		Conserved motif instances	12.7	193
		ChIP/Motif intersection	12.6	39
		Motifs without ChIP	12.7	154

Figure 2-12: Comparison of ChIP and comparative motif instances (at 40% confidence) for predicting regions and genes likely to be bound by a factor. Enhancer regions defined in Ernst et al. (2011). Regions within 2kb of a gene TSS are used as for assessing regulatory enrichment.

and Mef2) have comparable ability to find muscle genes (as defined in Tomancak et al., 2002), again with the intersection having considerably higher enrichment. Moreover, for the mesodermal repressor Snail, whose binding sites we expect to be avoided near mesodermal genes, we find that particularly true for comparative motif instances. Given that the motif instances have enrichments consistent with those of ChIP, a significant advantage beyond the low cost of the motif instances is their ability to suggest specific regulatory bases — a property we will take advantage of in Chapter 6.

While we see comparable enrichments for motifs and ChIP datasets in these functional defined regions, the sensitivity can differ dramatically between the two. For example, the factors shown in Figure 2-12 have many more peaks than the number of motif instances at 40% confidence. Consequently, when few comparative motif instances are available, ChIP may be more appropriate for identifying a broad range of the targets of a factor. However, these results suggest that comparative motif instances would perform comparably to ChIP when only a small number of confident instances are necessary.

2.7 Comparison to motif discovery

While *de novo* comparative motif discovery (Cliften et al., 2003; Kellis et al., 2003; Xie et al., 2005; Stark et al., 2007b) and instance prediction are related problems, motif discovery does not generally benefit from additional species as much as motif instance prediction does. Indeed, we have found that a small number of species appropriately placed is sufficient to statistically distinguish real motifs from fake ones (Figure 2-14). This is a consequence of motif discovery methods leveraging statistical over conservation across thousands of matches without needing to predict any individual instances. Moreover, we find that genome-wide conservation scores are highly correlated when computed both when using four species and the entire eutherian tree (Figure 2-15).

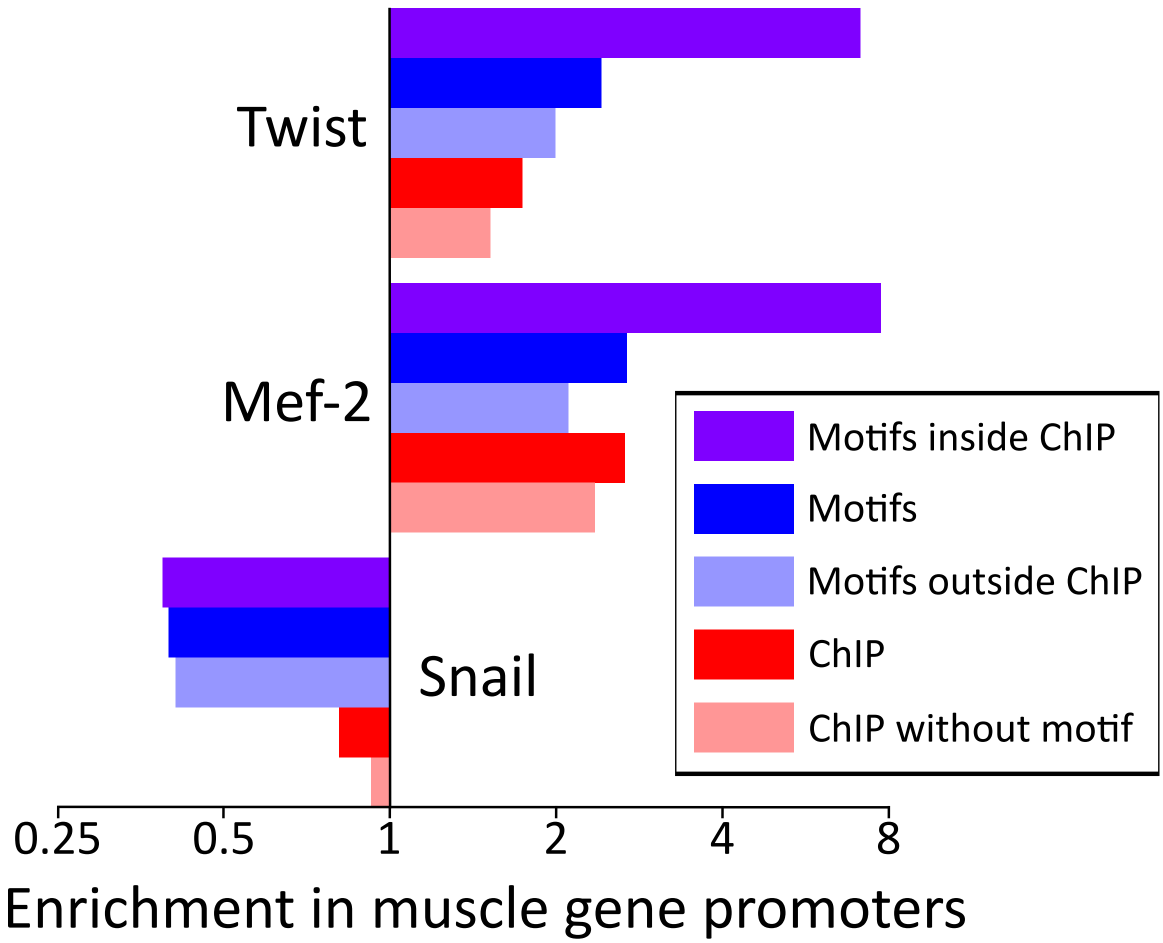


Figure 2-13: Enrichments of motifs and ChIP-chip in regions likely to be functional. We observe enrichments of mesodermal activators Twist and Mef-2 along with depletion of the repressor Snail within 2 kb upstream of genes expressed in embryonic muscles. We see that the motif instances (at 60% confidence) perform at least as well as ChIP and that the intersection of the two has dramatically higher enrichments. Except for Snail, all p-values are significant with $P < 5 \times 10^{-3}$.

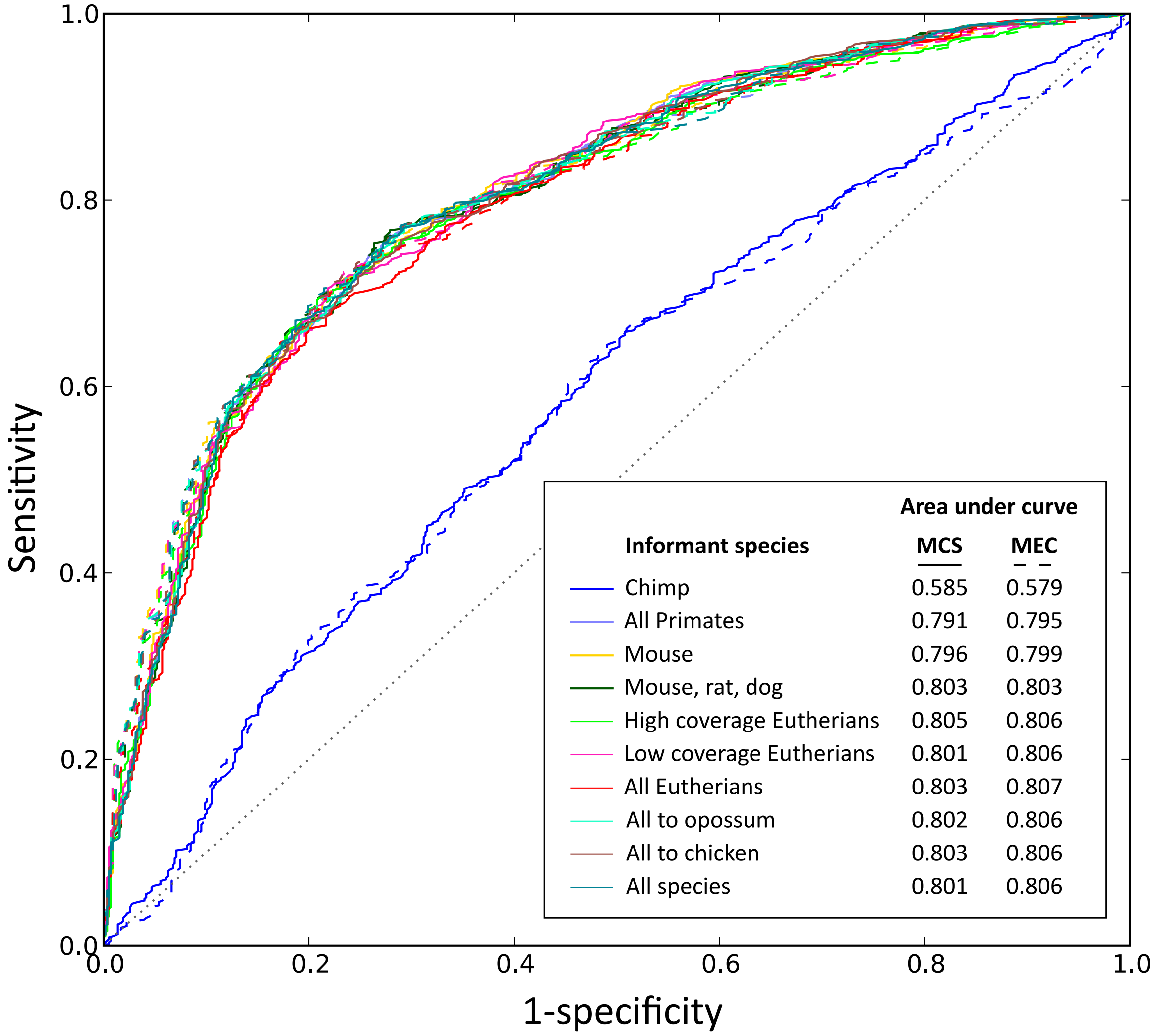


Figure 2-14: Distinguishing real from random motifs requires a limited number of species. ROC curves comparing different informant species subsets in predicting real motifs. Two methods are used to score motifs: MCS (Xie et al., 2005) in solid lines or MEC (Stark et al., 2007b) in dashed lines at varying BLS levels (without motif movement). All motifs with at least two shuffles (N = 577) in the known motif database were scored genome-wide to show a preference for being conserved at the optimal branch length score (in terms of AUC) for each species subset. Additionally, shuffles of these motifs were scored using the same criterion. Using only mouse, rat, and dog as informant species performs essentially identically to using the entire eutherian tree in separating the known and shuffled motifs. Indeed, even using just a single informant (mouse), has nearly equivalent performance. The two scoring schemes also distinguish between the two motif sets equally well. This demonstrates that at the number of instances and level of conservation seen for motifs in our database, motif discovery will likely not perform better when using motif conservation methods that employ a statistical conservation signal across the instances found genome-wide.

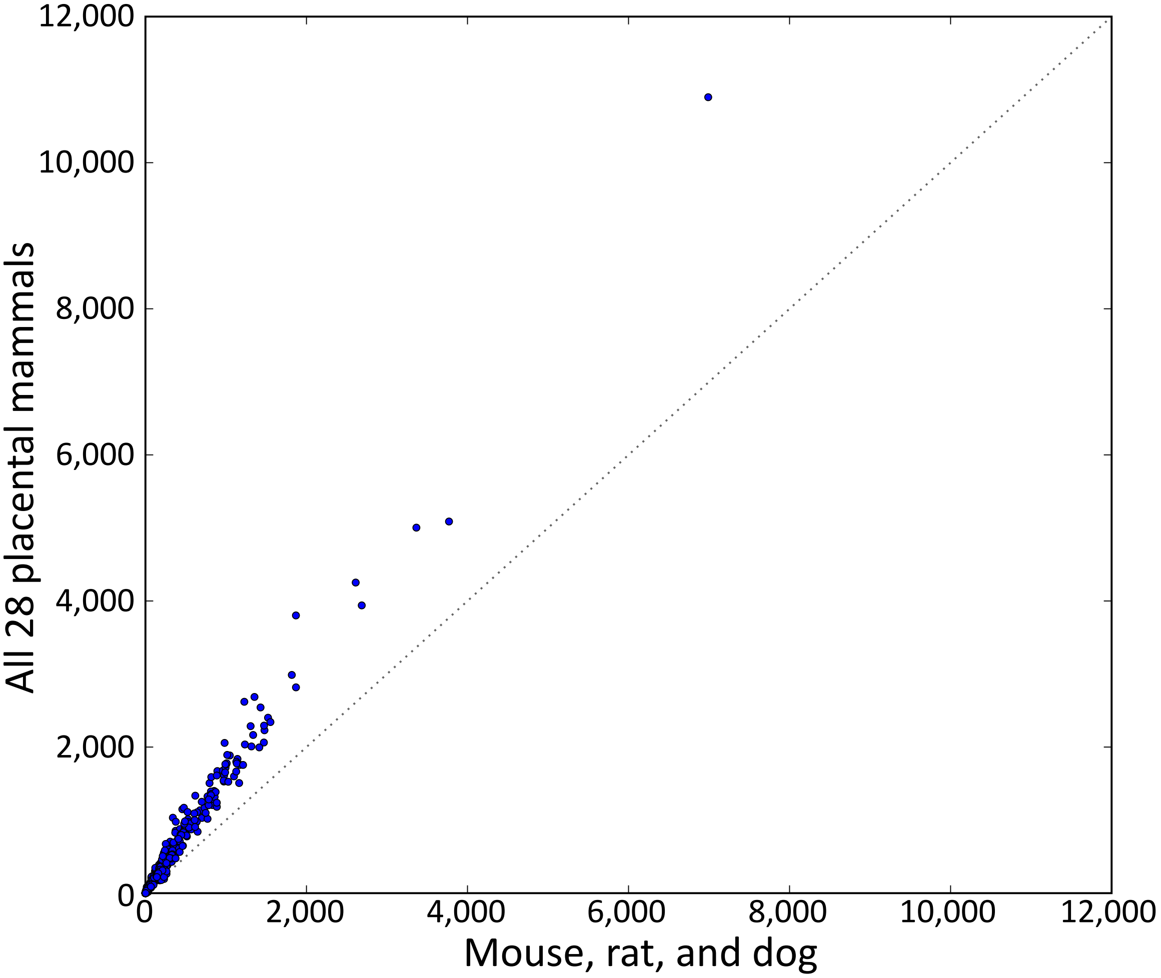


Figure 2-15: Motif conservation score (MCS; Kellis et al., 2003, 2004; Xie et al., 2005) is strongly correlated when using the entire eutherian tree or only mouse, rat and dog as informant species. A correlation of 0.99 is seen between the MCS scores on known motifs computed using all eutherian informant species and when only using mouse, rat, and dog as informants (r=0.99). MCS is computed as the negative log of the hypergeometric p-value of the conservation rate of the motif compared to the controls for that motif. This analysis does not preclude the possibility of the existence of motifs specific to a given clade (e.g., the primates) that are yet to be described.

2.8 Conclusion and future directions

In this chapter we showed that comparative genomics is a powerful tool in predicting likely functional motif instances. We present an algorithm for finding comparative motif instances that is robust to low quality genomes, alignment difficulties, and common evolutionary events. We also show that these motif instances are correlated with *in vivo* binding and have competitive enrichment with ChIP for identifying genes likely to be functional for a given TF. The main biological contribution is the prediction of specific target genes for hundreds of factors in human and fly, each with a specified confidence level.

The methods presented in this chapter have been adapted and extended by others. Xie et al. (2009) modify BLS with a Bayesian approach that permits partial matches to a species, producing a Bayesian BLS. Friedman et al. (2009) extend our work in a different direction customizing it for use with miRNA binding sites and show that correcting for the variable nature of conservation in the genome can lead to increased signal and more confidently predicted motif instances. Due to the rigorous statistical approach and control motifs, method improvements do not invalidate the stated specificity of predicted motif instances, but rather at best can increase the number that can be predicted.

In addition to these changes, a number of extensions could improve the prediction of motif instances. Much of the power of BLS comes from its lack of making specific model assumptions (such as a strict model of evolution). However, a consequence of this is that a strongly non-phylogenetically consistent collection of matching species (such as only a distant informant), may obtain a high BLS score. Additional research is required to determine a good way to create a more model-based approach while retaining the robustness of BLS (for example, but incorporating the number of matching species).

Chapter 3

Systematic characterization of motifs in transcription factor binding experiments

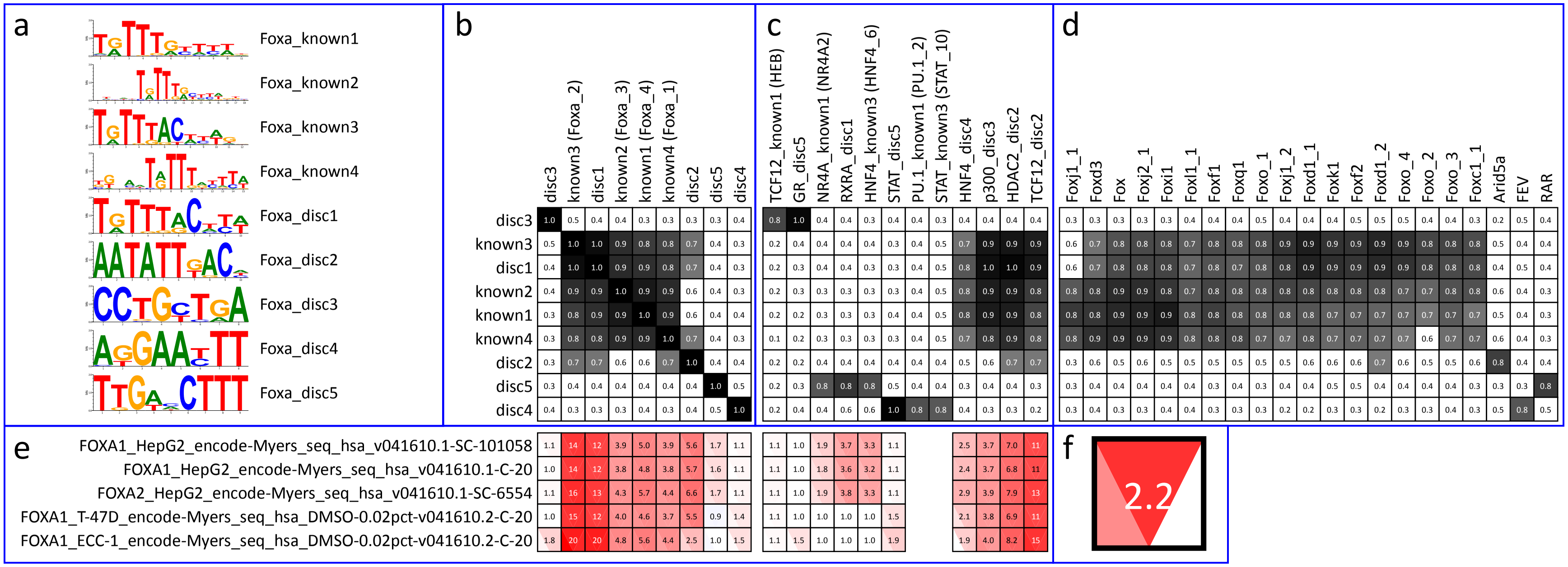
This chapter presents a systematic analysis of the motif content of hundreds of human ChIP-seq datasets carried out on transcription factors. It represents work done as part of the ENCODE consortium, from which the data is derived. I took the peak data from ENCODE and developed and applied the motif discovery pipeline, and did the subsequent analysis presented in this chapter.

3.1 Introduction

Chromatin immunoprecipitation (ChIP; Solomon et al., 1988) followed by hybridization to an array (ChIP-chip; Ren et al., 2000; Iyer et al., 2001) or sequencing (ChIP-seq; Robertson et al., 2007) enable the genome-wide identification of the binding sites of transcription factors present in a given sample. As these technologies have matured, their use has become increasingly widespread. The regions called as peaks by these experimental techniques can be as small as 300 bp for ChIP-chip (Qi et al., 2006) and 50 bp for ChIP-seq (Guo et al., 2010), depending on the experimental design and algorithmic processing of the raw data. However, many applications require a deeper understanding of the regulatory code which necessitates (1) identifying the relevant motif(s) for each factor, (2) understanding the interplay between them, and (3) predicting their instances. To address this need, we have performed a systematic motif-centric analysis of the human TF binding profiles produced by ENCODE.

Predicting the motifs responsible for TF binding is an active area of research. It is complicated by computational difficulties in finding enriched patterns (Tompa et al., 2005), the binding of many factors in close proximity (Moorman et al., 2006), the complex and varying composition of the genome (Lander et al., 2001), and the prevalence of peaks without a clear motif match (Li et al., 2008). One popular approach to deal some of these problems is to use multiple *de novo* motif

Figure 3-1 *(on the next page)*: Output shown for the Foxa factor group, a relatively representative example. (a) The known and discovered motifs for the factor group, drawn with WebLogo (Crooks et al., 2004). Because the original orientation is arbitrary they are flipped as necessary to increase the similarity of the displayed orientations. (b-d) Similarity between the motifs for this factor group and the motifs: (b) for this factor group, (c) for all other factor groups, and (d) other known motifs from our database that were not used as a known motif for any factor group. For (c,d) the motif must match at least one motif for this factor with similarity ≥ 0.75. Similarities are shown in black/white scale with gray starting at 0.65. Names on top in parentheses match known motif names used in Lindblad-Toh et al. (2011). (e) Enrichments of motifs in (b-c) for the datasets for this factor group. Datasets are named indicating the factor, cell type/stage, lab code, followed by values to differentiate datasets and make them unique. Red is used for enrichment over 1, blue is used for enrichment below 1 (depletion; which is rare in this study). Enrichments are not shown for motifs when control motifs could not be generated or there wasn't sufficient information content to match at a 4^{-8} p-value. (f) Magnified enrichment heatmap value. The three triangles represent the background regions used for enrichment; the top triangle is all regions, the left triangle is those within 2kb of an annotated TSS and the right triangle is those outside the 2kb window (the regions used in the left and right triangles partition that of the top triangle). The number shown is the enrichment in the inclusive background. Here we see an apparent contradiction: a higher enrichment for the union than the parts. This occurs because the higher counts permit a smaller confidence interval around the ratios used to compute the enrichment. Heatmaps are ordered using hierarchical clustering followed by optimal leaf ordering (Bar-Joseph et al., 2001).



	disc3	known3 (Foxa_2)	disc1	known2 (Foxa_3)	known1 (Foxa_4)	known4 (Foxa_1)	disc2	disc5	disc4
disc3	1.0	0.5	0.4	0.4	0.3	0.3	0.3	0.3	0.4
known3	0.5	1.0	1.0	0.9	0.8	0.8	0.7	0.4	0.3
disc1	0.4	1.0	1.0	0.9	0.9	0.8	0.7	0.4	0.3
known2	0.4	0.9	0.9	1.0	0.9	0.9	0.6	0.4	0.3
known1	0.3	0.8	0.9	0.9	1.0	0.9	0.6	0.4	0.4
known4	0.3	0.8	0.8	0.9	0.9	1.0	0.7	0.4	0.3
disc2	0.3	0.7	0.7	0.6	0.6	0.7	1.0	0.4	0.4
disc5	0.3	0.4	0.4	0.4	0.4	0.4	0.4	1.0	0.5
disc4	0.4	0.3	0.3	0.3	0.4	0.3	0.4	0.5	1.0

	TCF12_known1 (HEB)	GR_disc5	NR4A_known1 (NR4A2)	RXRA_disc1	HNF4_known3 (HNF4_6)	STAT_disc5	PU.1_known1 (PU.1_2)	STAT_known3 (STAT_10)	HNF4_disc4	p300_disc3	HDAC2_disc2	TCF12_disc2
disc3	0.8	1.0	0.4	0.4	0.3	0.4	0.4	0.3	0.4	0.4	0.4	0.4
known3	0.2	0.4	0.4	0.4	0.4	0.3	0.4	0.5	0.7	0.9	0.9	0.9
disc1	0.2	0.3	0.4	0.4	0.3	0.3	0.4	0.5	0.8	1.0	1.0	0.9
known2	0.2	0.3	0.3	0.3	0.3	0.3	0.4	0.5	0.8	0.9	0.9	0.8
known1	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.5	0.8	0.9	0.9	0.8
known4	0.1	0.2	0.3	0.3	0.3	0.3	0.4	0.5	0.7	0.8	0.9	0.8
disc2	0.2	0.2	0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.6	0.7	0.7
disc5	0.2	0.3	0.8	0.8	0.8	0.5	0.4	0.4	0.3	0.4	0.4	0.3
disc4	0.2	0.4	0.4	0.6	0.6	1.0	0.8	0.8	0.4	0.3	0.3	0.2

	Foxj1_1	Foxd3	Fox	Foxj2_1	Foxi1	Foxl1_1	Foxf1	Foxq1	Foxo_1	Foxj1_2	Foxd1_1	Foxk1	Foxf2	Foxd1_2	Foxo_4	Foxo_2	Foxo_3	Foxc1_1	Arid5a	FEV	RAR
disc3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.5	0.4	0.4	0.4	0.5	0.5	0.5	0.4	0.4	0.3	0.2	0.5	0.4
known3	0.6	0.7	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.5	0.4	0.4
disc1	0.6	0.7	0.8	0.8	0.8	0.7	0.8	0.7	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.4	0.5	0.4
known2	0.8	0.8	0.9	0.9	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.7	0.8	0.8	0.6	0.4	0.3
known1	0.8	0.8	0.9	0.9	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.5	0.4	0.3
known4	0.8	0.9	0.9	0.9	0.9	0.8	0.8	0.8	0.7	0.7	0.8	0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.6	0.4	0.3
disc2	0.3	0.6	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.7	0.6	0.5	0.5	0.6	0.8	0.4	0.4
disc5	0.3	0.4	0.4	0.4	0.4	0.4	0.3	0.4	0.3	0.4	0.4	0.5	0.5	0.4	0.4	0.4	0.4	0.3	0.3	0.3	0.8
disc4	0.3	0.4	0.3	0.3	0.4	0.4	0.3	0.3	0.3	0.2	0.2	0.3	0.3	0.2	0.3	0.3	0.4	0.3	0.5	0.8	0.5

FOXA1_HepG2_encode-Myers_seq_hsa_v041610.1-SC-101058	1.1	14	12	3.9	5.0	3.9	5.6	1.7	1.1	1.1	1.0	1.9	3.7	3.3	1.1			2.5	3.7	7.0	11
FOXA1_HepG2_encode-Myers_seq_hsa_v041610.1-C-20	1.0	14	12	3.8	4.8	3.8	5.7	1.6	1.1	1.1	1.0	1.8	3.6	3.2	1.1			2.4	3.7	6.8	11
FOXA2_HepG2_encode-Myers_seq_hsa_v041610.1-SC-6554	1.1	16	13	4.3	5.7	4.4	6.6	1.7	1.1	1.1	1.0	1.9	3.8	3.3	1.1			2.9	3.9	7.9	13
FOXA1_T-47D_encode-Myers_seq_hsa_DMSO-0.02pct-v041610.2-C-20	1.0	15	12	4.0	4.6	3.7	5.5	0.9	1.4	1.1	1.0	1.0	1.0	1.0	1.5			2.1	3.8	6.9	11
FOXA1_ECC-1_encode-Myers_seq_hsa_DMSO-0.02pct-v041610.2-C-20	1.8	20	20	4.8	5.6	4.4	2.5	1.0	1.5	1.1	1.5	1.0	1.0	1.0	1.9			1.9	4.0	8.2	15

discovery tools (e.g., Che et al., 2005; Romer et al., 2007; Sun et al., 2010). This is strategy that we take, which has the important feature of producing diverse list of motifs for each factor.

The main contribution of this chapter is the systematic application of motif discovery to hundreds of datasets along with the integration of motifs collected from the literature. These motifs along with carefully performed enrichment analysis in the relevant datasets are available at http://compbio.mit.edu/motif-disc/human. The interface (Figure 3-1) allows for browsing each factor group individually, displaying the known (literature) and discovered motifs and their similarities to each other and to motifs of other factor groups. For each motif the enrichment of its instances is indicated in each of the relevant datasets.

Together these features permit readily identifying many of the results we present here. Along with these statistics, the underlying data files for this analysis, including the motif matrices and their genome-wide matches, are provided. The breath of datasets enables systematic comparisons and analyzes that are not possible when only one or a few factors are studied in isolation. The remainder of this chapter details our *de novo* discovery pipeline and examples of the type of analysis it enables.

3.2 Methods

Human motifs are collected from various large scale databases, as described in Section 1.5.1. Enrichments are computed as described in Section 5.2, using a Wilson's interval corrected ratio of ratios.

3.2.1 Comparing motifs

Motif similarity is defined as the maximal Pearson's correlation of the PWM made into a 4xN vector across all offsets and both orientations, padding unmatched positions with N's (Pietrokovski, 1996). Motifs are considered a match if they have a similarity of at least 0.75; weak matches or variants are determined through manual inspection. We found that the cutoff of 0.75 to be the limits of similarity shared between variants of motifs for the same factor, although it is ultimately arbitrary.

3.2.2 Processing and naming of experimental datasets

Human ENCODE (Consortium, 2011a) protein binding datasets (excluding Pol2 and Pol3) were taken from the January, 2011 freeze. They were processed uniformly using SPP (Kharchenko et al., 2008) as described in Kundaje et al. (in review). To avoid potentially confounding issues, we excluded from our analysis: (1) the Y and mitochondrial chromosomes; (2) the hg19 rmsk and simple repeat tracks from UCSC (created on April 27, 2009; Kent et al., 2002); and (3) protein coding regions, exons for non-coding genes, and $3'$ UTRs taken from Gencode v4 Harrow et al. (in review).

3.2.3 Performing *de novo* motif discovery

Peaks were randomly partitioned into two datasets for the purpose of separating discovery and enrichment and limit over-fitting. The top 250 peaks for the first partition were used in motif discovery (high intensity peaks generally have better enrichment for motifs; MacArthur et al., 2009). Five tools were run independently run on each dataset: AlignACE (v4.0 with default parameters; Hughes et al., 2000), MDscan (v2004 with default parameters; Liu et al., 2002), MEME (v3.5.7 with a maximum of 10 iterations and -maxw 10, 15, and 25 for the 3 motifs; Bailey and Elkan, 1994), Weeder (v1.4.2 with option large; Pavesi et al., 2001), and Trawler (v1.2 with 250 random intergenic blocks for background; Ettwiller et al., 2007). Any motifs beyond the top three for any method on one dataset were discarded.

3.2.4 Selecting and ordering discovered motifs

For each factor group we order the motifs from all discovery programs by the enrichment in the second random partition that was held out for dataset where they were discovered, where for this step only we randomly select 10% of the background regions to reduce the amount of computation. We then select discovered motifs for each factor by this rank order discarding any motif that matches a previous one with similarity greater than 0.75. Because this enrichment is compared to background regions, patterns that are simply common in the genome but not more common in the bound regions are not selected. Moreover, because the enrichment must be greater than 1 (i.e. no enrichment), and because the enrichment procedure we employ uses confidence intervals (Section 5.2), all selected motifs are effectively statistically significantly enriched in the dataset for which they are taken.

We supplement these discovered motifs with the known motifs from the literature (described above) and rematch the motifs to the entire background regions and produce comparable enrichments for all datasets.

3.3 Resource description

Our goal with this resource was to (1) produce a collection of varied, enriched motifs for each factor; (2) avoid repetitive, weakly enriched motifs that do not capture real biology of the factor; and (3) avoid variants of the same motif. With this in mind, motif discovery is conducted separately on each dataset using five motif discovery tools (Figure 3-2) and each factor and all its datasets are manually placed into 'factor groups' on the basis of known motifs and homology. Known motifs from the literature and the top 10 most enriched discovered motifs (disallowing redundancy) are collected for each factor group and named as factor-group_known# for known motifs and factor-group_disc# for discovered motifs. Known motifs are ordered arbitrarily, whereas the discovered motifs are ordered in descending order of the enrichment value that was used to choose them.

One imperfect way to judge this procedure is to compare the discovered motifs against those previously identified, mainly *in vitro* motifs. Recovery of known motifs varies significantly by method, but taking the most enriched motif (our pipeline) is competitive with the best single method (Figure 3-3). However, the main advantage of our pipeline is that it produces several dissimilar motifs that can be used to analyze properties of the factors and datasets and thereby make connections between them.

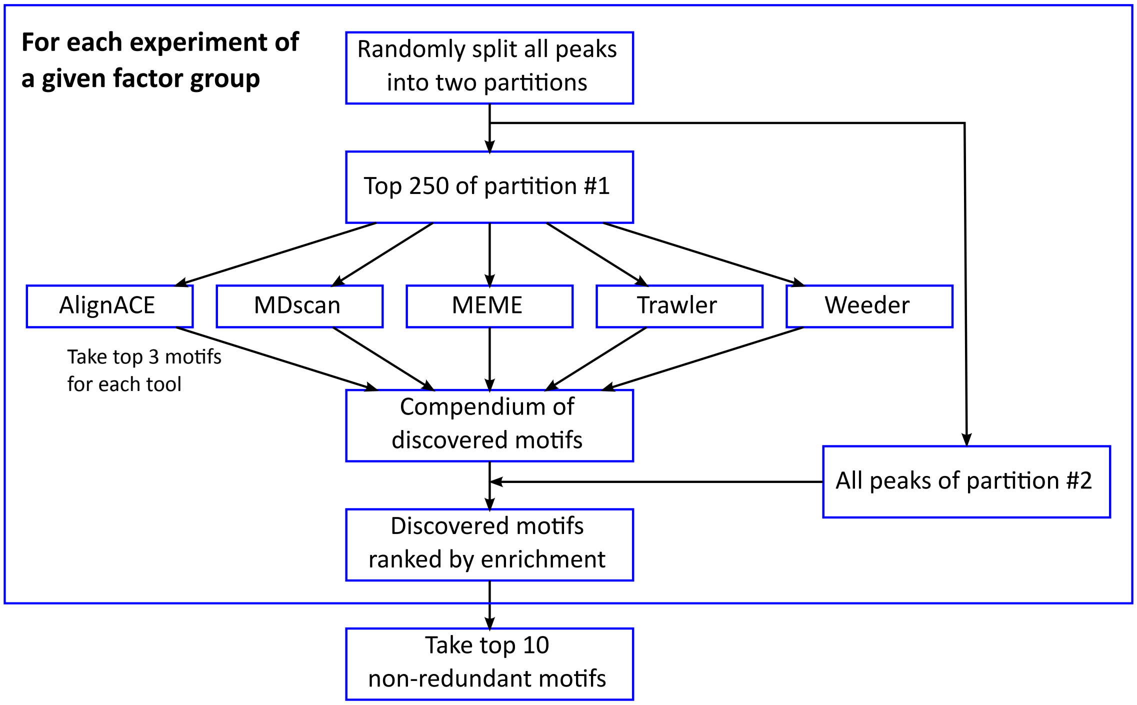

For each experiment of a given factor group
Randomly split all peaks into two partitions
Top 250 of partition #1
AlignACE
MDscan
MEME
Trawler
Weeder
Take top 3 motifs for each tool
Compendium of discovered motifs
All peaks of partition #2
Discovered motifs ranked by enrichment
Take top 10 non-redundant motifs

Figure 3-2: Outline of motif discovery pipeline.

	Datasets	Factors	Factor groups	Factor groups with literature motif
Human	427	125	84	53

Table 3-1: Statistics on the datasets used for the analysis.

Ultimately, only 12 of the 84 factor groups had no enriched discovered motif. Nine of these TF groups contain factors with no known DNA binding domain

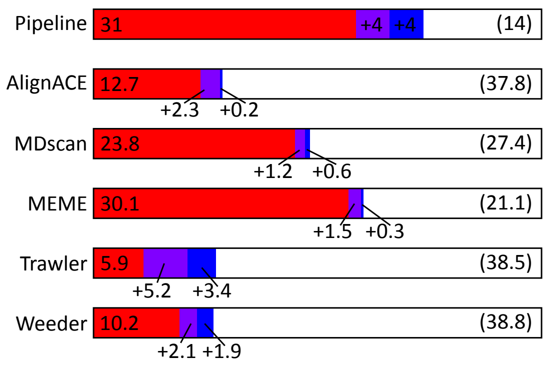


Figure 3-3: Performance of discovery in terms of number of factor groups for which the known motif was recovered. A motif is considered a match if it matches any of the known motifs for a factor group (see Methods for details on how matches are computed). The number of additional factors that have a match is shown with each additional motif (only 3 motifs are taken from each individual method, whereas we have up to 10 for the pipeline). The number of factor groups with no motif match is shown in parenthesis. When multiple datasets exist for a factor group, the fraction that match is used in computing its contribution for computing the performance of the individual tools.

(BRF, CtBP2, GCN5, HDAC8, NELFe, SPT20, SUZ12, WHIP, and XRCC4) as indicated by UniProt (Bairoch, 2004) while the remaining three (NR4A, ZNF274, and ZZZ3) do. Moreover, while these twelve TF groups cover 22/427 datasets (5%), they account for 12 of the 56 (21%) datasets flagged as unreliable based on various quality metrics (all datasets for BRF, GCN5, NELFe, NR4A, SPT20, and ZZZ3; see Kundaje et al., in review), suggesting this lack of motifs may not reflect the biology of some of these factors. Of these factors, only NR4A has a previously identified known motif and that motif is not enriched in the dataset.

Through manual inspection, we identified 46 discovered motifs that are either low-complexity (e.g., dinucleotide repeats) or consistently have weak enrichment (<2) and do not match known motifs. These represent 16% of the 293 discovered motifs are likely either due to slight biases in the discovery pipeline or due to real but relatively weak specificity for the factor. We will generally exclude these motifs from further discussion, however they remain available on the website for any relevant analysis: AP-1_disc10, ATF3_disc3-4, BDP1_disc3, BHLHE40_disc2, CHD2_disc2-3, E2F_disc7-8, EBF_disc2, ELF1_disc2-3, ERalpha-a_disc4, Egr-1_disc6, Ets_disc9, GATA_disc5, GR_disc6, HDAC2_disc6, Hey1_disc2, Myc_disc9-10, Nrf_disc3, NRSF_disc4-5, p300_disc8-9, PAX_disc5, PU.1_disc3, Pou2f2_disc2, Rad21_disc5-8, SRF_disc2, STAT_disc6-7, Sin3Ak-20_disc5-7, TATA_disc10, TCF12_disc5-6, YY1_disc3-5, and Znf143_disc4.

Further, while our pipeline inherently avoids variants of the same discovered motif, because correlation does not capture all similarity properties between motifs, some variants were still found. This most frequently is seen for longer motifs that can be broken up into recognizable but globally dissimilar patterns. Like with the low complexity motifs, we ignore these motifs that have an apparent similarity to the known motif for the factor but for which a better matching and enriched motif is also found: CHD2_disc3, CTCF_disc2-7,10, E2F_disc5, Egr-1_disc7, ERalpha-a_disc3, Ets_disc4, Foxa_disc2, GATA_disc4, HNF4_2-3, Irf_disc6, Myc_disc6-8, Nrf1_disc3, NF-E2_disc4, NRSF_disc2-3,6-7,10, Rad21_disc2,4,9-10, STAT_disc5, Sin3Ak-20_disc3-4, TCF12_disc3, and ZBTB7A_disc2. Some variants also existed amongst the discovered motifs, and they are indicated as “weak matches” when discussed.

3.4 Biological results and example resource applications

Here we present an incomplete analysis of the biological insights an examination of our resource enables. In the interest of clarity, most descriptions of TFs will be omitted, but may be found along with further references at RefSeq (Pruitt and Maglott, 2001) and Entrez (Maglott et al., 2007).

3.4.1 Recovery of the known specificity for TFs

After removing variants and low complexity motifs, 11 of the 53 factor groups with a known motif (BHLHE40, EBF, Maf, NF-Y, NF-kappaB, NRSF, Pou2f2, SRF, YY1, ZBTB7A, and ZEB1) had only one discovered motif found which matched the corresponding known motif for that factor group. The fact that most factors are enriched for multiple, distinct motifs highlights the complexity of TF binding and the importance of interaction between factors.

Much more common than the case of finding a single motif matching a previously characterized motif for a factor is finding several motifs one of which matches the known. Indeed, this is what we see for 20 factor groups (CEBPB, CTCF, ERalpha-a, Egr-1, Foxa, GATA, GR, HNF4, Mef2, Myc, NF-E2, Nrf1, PRDM1, PU.1, Pax-5, Pou5f1, RFX5, RXRA, STAT, and TCF12) where the most enriched discovered motif (disc1) matches a known motif for that factor group. Moreover, for 4 factor groups (AP-2, Ets, Mxi1, and TCF4) the second discovered motif (disc2) matches the known and for an additional 4 groups (AP-1, E2F, Irf, and SP1) the third discovered motif (disc3) matches. Consequently, for each of these factor groups several additional motifs were discovered that appear to have meaningful matches to either known motifs for other factors or to novel non-repetitive motifs. In the next section will describe the additional motifs we find for these factors, which in many cases are factors known to interact (either cooperatively or competitively) and in others make predictions for interacting partners suitable for additional investigation.

For several factor groups (e.g., Hsf, Nanog, Pbx3, SREBP, and TAL1) the known motif is not found at all. Frequently this is because the known motif itself is not enriched and may not accurately capture the specificity of the factor *in vivo*. For example, the “known” p300 motif from Transfac was likely built on a very specific bound region of p300 and would not accurately capture its binding in all cell types where it interacts with a variety of factors and has no DNA binding domain of its own (we avoided removing such motifs to prevent bias in the dataset). Likewise, the known ZBTB33 motif is not enriched at all in the bound regions and unsurprisingly we do not discovery a motif that matches it.

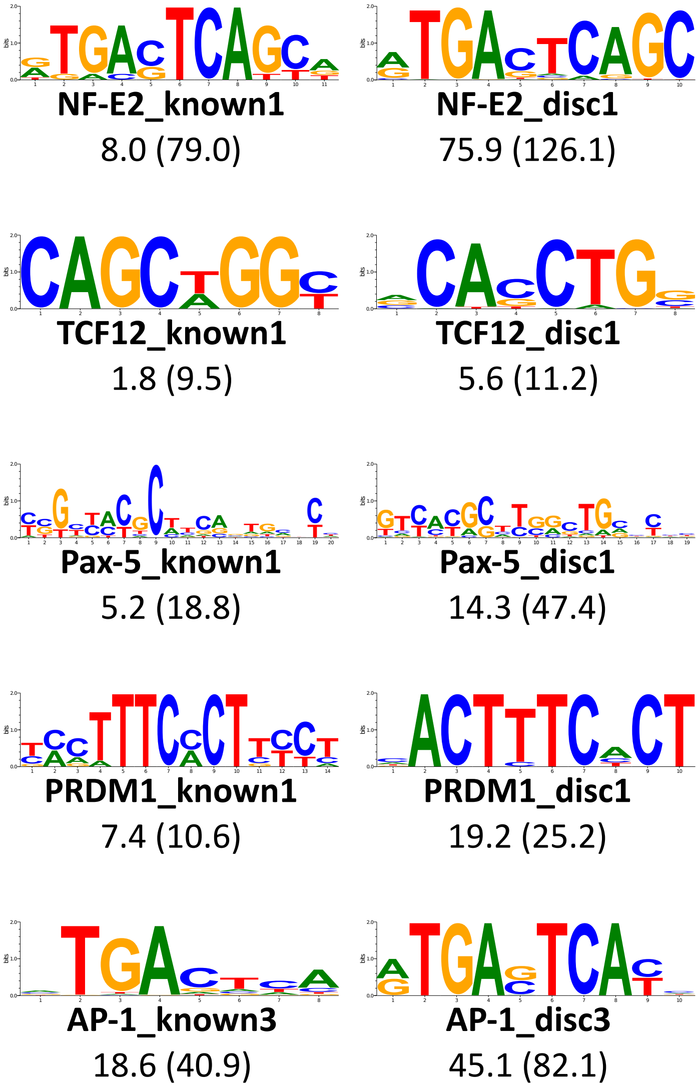
While some known motifs were problematic, we largely found our database of known motifs to be relatively comprehensive and had difficulty finding matches to novel motifs outside it. An exception is ZNF263_disc1 which does not match a motif in our database, but does roughly match the specificity for ZNF263 indicated in Frietze et al. (2010) despite only having weak enrichment (1.8-fold).

While the motifs that match each other (either known or discovered) generally have similar enrichments, in some cases we find substantially higher enrichment for some motif variants over others (Figure 3-4). For example PRDM1_disc1 matches the known PRDM1 motif but is 19-fold enriched, compared to 7-fold for the most enriched known PRDM1 motif, even though both are similar. Sometimes the more enriched motif is for a more distant family member: for example, for ELF1 (an Ets factor) we did not find the corresponding known motif, however, ELF1_disc1 matches other Ets motifs and is about twice as enriched as the known motif. Known motifs often show a broad range in enrichment: Mef2 has 6 motifs described in Transfac, with an enrichment differential of as much as 4-fold consistently across datasets. Our resource provides a principled way to choose amongst variants of a motif.

Differences in enrichment of the known motif for a factor across datasets for the same factor was frequently seen. For example, the known CTCF motif is enriched in datasets in a range of 23- to 62-fold on identically processed data. These examples are suggestive of problematic datasets, but could also be a consequence of the different biology of these cell types.

Generally the discovered motifs we find are not able to distinguish between families of factors. An exception is ERalpha vs. ERRA (which we place into the ERalpha-a factor group). The ERR-alpha motif is two facing ERalpha-a sites. We recover these two motifs as ERalpha-a_disc1 (matching the known ERalpha-a motif) and ERalpha-a_disc2 (matching the known ERRA). We find enrichments that mirror this relationship: ERRA datasets are more enriched for disc, while ERRalpha-a datasets are more enriched for disc1.

Twenty factor groups had no known motif but now have discovered enriched motifs (BAF155, BATF, BCL, BDP1, CCNT2, CHD2, CTCFL, HDAC2, HEY1, HMGN3, KAP1, Rad21, SETDB1, SIRT6, SMC3, SP2, Sin3A, THAP1, TR4, and



NF-E2_known1
8.0 (79.0)

NF-E2_disc1
75.9 (126.1)

TCF12_known1
1.8 (9.5)

TCF12_disc1
5.6 (11.2)

Pax-5_known1
5.2 (18.8)

Pax-5_disc1
14.3 (47.4)

PRDM1_known1
7.4 (10.6)

PRDM1_disc1
19.2 (25.2)

AP-1_known3
18.6 (40.9)

AP-1_disc3
45.1 (82.1)

Figure 3-4: Comparison of known versus discovered motifs. Displayed is the known and discovered motif with the maximum enrichment across all datasets for a factor group. Only the discovered motifs that match a known motif for a factor group are considered. The maximum enrichment is indicated for each factor and, in parenthesis, the "raw" enrichment for the same dataset without the use of the shuffle motifs for correction.

ZNF263). These discovered motifs may represent the direct or indirect (e.g., through cofactors) DNA binding specificity.

3.4.2 Shared motifs suggest interacting relationships

One of the most striking features of this analysis is that most factors have motifs for other factors enriched in their binding sites. This may occur due to: (1) the cooperative binding of the two factors to the same locations; (2) an interfering relationship between factors where one binds near the other to prevent binding; (3) high similarity in binding specificity; (4) the two factors functioning on a similar set of genes (e.g., ones specific to one tissue), without directly interacting; or (5) the factors binding to similar genomic regions (e.g., near genes). Our analysis doesn't directly rule out any of these possibilities, however, (3) is generally verifiable using our motif similarity metrics and (5) can be examined by inspecting only the TSS-proximal enrichment.

The motif most enriched in multiple datasets was the TPA DNA response element (TRE; TGA[C/G]TCA), which is recognized by the AP-1 transcription factor when it is formed by Fos/Jun dimers (Karin et al., 1997) and other factors including Maf and NF-E2. The enrichment of the TRE in a dataset is often stronger than that of even the known *in vitro* sequence specificity and may arise from a number of phenomena, including (1) a cooperatively interaction with AP-1, (2) competition with AP-1 for the same binding sites, leading to a potentially repressive role for the TF, or (3) reuse of binding sites due to, for example, accessibility of chromatin. We find a motif matching the TRE motif for 20 factor groups (AP-1_disc3, AP-2_disc1, BAF155_disc1, BATF_disc1, BCL_disc2, CTCF_disc8-9, GATA_disc2, GR_disc2, HMGN3_disc1, Irf_disc2, KAP1_disc1, Maf_disc1, Mef2_disc3, Myc_disc3, NF-E2_disc1, p300_disc1, PRDM1_disc2, RXRA_disc3, STAT_disc2, and TCF4_disc1).

We found that the enrichment of the TRE to be particularly notable for a few factors. GATA and AP-1 have known cooperative binding (Kawana et al., 1995).

Factor	Matching discovered motifs
CTCF	Rad21_disc1 (Wendt et al., 2008); SMC3_disc1 (Rubio et al., 2008); CTCFL_disc1 (Jelinic et al., 2006)
CEBPB	STAT_disc4 (Choi et al., 2007)
Ets	GATA_disc3 (Rothbächer et al., 2007); Mef2_disc2 (Taylor et al., 1997); TR4_disc1 (O'Geen et al., 2010)
GATA	TAL1_disc1 (Kappel et al., 2000; Mouthon et al., 1993)
Mrg	Pbx3_disc2 (Bischof et al., 1998)
Myb	Ets_disc8 (Dudek et al., 1992)
Myc	Sin3Ak-20_disc2 (Nascimento et al., 2011)
NF-Y	Irf_disc1 (Li-Weber et al., 1994); CEBP_disc2 (Yu et al., 1995); E2F_disc2 (Caretti et al., 2003); RFX5_disc2 (Villard et al., 2000); SP1_disc1-2 (Roder et al., 1999)
NRSF	Sin3Ak-20_disc1 (Huang et al., 1999)
PU.1	Irf_disc5 (Scott et al., 1994)
Pou5f1	Nanog_disc2 (Looijenga et al., 2003; Loh et al., 2006)
TRE	BAF155_disc1 (Ito et al., 2001); GATA_disc2 (Kawana et al., 1995); TCF4_disc1 (Nateri et al., 2005); STAT_disc2 (Ivanov et al., 2001)
YY1	THAP1_disc1 (Mazars et al., 2010; Yu et al., 2010)

Table 3-2: Selected shared motifs with literature support. Shown are the motifs that match a known motif for the indicated factor along with relevant citations. Details in the text.

BAF155, BAF170, Brg1, and Ini1 (which we place into the BAF155 factor group) are members of the SWI/SNF chromatin remodeling complex (Wang et al., 1996) which is necessity for proper regulation by Fos/Jun dimers (Ito et al., 2001); and TCF4_disc1, which matches the TRE, is more enriched than the known TCF4 motif (TCF4_disc2) in only the TCF4 colorectal cancer cell line HCF-116 dataset, consistent with the known interaction of Jun and TCF4 during intestinal cancer development (Nateri et al., 2005).

AP-1 also binds to the cAMP response elment (CRE; TGACGTCA) when the dimer is formed by ATF3/Jun (Karin et al., 1997) and this is the motif we find as AP-1_disc1. However, AP-1_disc3 (which matches the TRE) is the most enriched motif in Fos datasets. Interestingly, ATF3_disc1 is not the CRE, but rather the E-box (see below). We do, however, find a variant of the CRE (with additional specificity) as ATF3_disc2. The most enriched discovered motif for E2F, E2F_disc1 also matches the CRE and is highly enriched in all datasets.

Myc is a critical regulator which recognizes the E-box sequence. To aid in comparisons, we include Max, which forms complexes with Myc, and USF1/2, which also recognizes the E-box sequence, in the Myc factor group. We find multiple motifs enriched in Myc binding sites, highlighting the multifunctional role Myc and the other E-box recognizing proteins play. We found a version of the E-box with additional specificity (Myc_disc1) was highly enriched in USF1/2 bound regions (max 98-fold for USF2 vs. <9-fold enrichment for Myc/Max). This motif was also more enriched than the known E-box motifs in USF datasets, including compared to motifs originally identified for USF. We find a second, less specific E-box motif (Myc_disc2), which shows more even enrichment across factors. Morever, Myc_disc4 matches RFX5 and is enriched particularly for Max bound regions in H1-hESC and GM12878, and Myc_disc5 matches the CEBPB known motif and is enriched in Myc regions bound in unstimulated K562 cells. Mxi1, which was not included in the Myc factor group although it does interact with Max to bind to Myc-Max sites (Zervos et al., 1993), has Mxi1_disc1 which matches RFX5 in both the K562 and HeLa-S3 cell lines. We also find discovered motifs of other factors matching the E-box, including Sin3Ak-20_disc2 (discussed below), NF-E2_disc2-3, and SIRT6_disc1. It is notable that while SIRT6 is a chromatin-associated protein without a known DNA binding domain (Mostoslavsky et al., 2006), the only discovered motif matches the E-box (with 16-fold enrichment in SIRT6 bound regions) suggesting that Myc or another E-box recognizing factor may play an important, but indirect chromatin-related role.

Motif enrichment is able to predict both positive and negative interactions for the same factor. For example, Sin3A, a co-repressor known to interact with a number of proteins, has discovered motifs matching NRSF (Sin3Ak-20_disc1 and more weakly disc3-4) and Myc (Sin3Ak-20_disc2). These are consistent with Sin3A's known involvement in repression by NRSF (Huang et al., 1999) and Sin3A being a known antagonist for Myc (Nascimento et al., 2011).

We analyzed six Irf family datasets: Irf1 binding in K562 cells stimulated by IFNa (viral innate response) or IFNg (viral, bacterial, and tumor control); Irf3 in HepG2, GM12878, and HeLa-S3; and Irf4 in GM12878. The most strongly enriched motif (Irf_disc1, matching NF-Y) is very highly enriched (>20-fold) for all three Irf3 datasets and Irf1 in K562 under IFNg stimulation. This suggests that binding of Irf to NF-Y sites occurs only under specific conditions and by only some Irf members and potentially expands on the previously documented interaction of NF-Y and Irf2 at a single promoter (Li-Weber et al., 1994). Irf_disc4, which matches SP1, is enriched in the same cell types, albeit at much lower levels. Irf_disc3, which matches the known Irf consensus, shows weak-to-no enrichment in these datasets, but shows an enrichment of 8.8-fold for Irf1 bound regions in K562 cells under IFNa stimulation and 3.1-fold enrichment for Irf4 bound regions in GM12878. Irf_disc2, which matches the TRE, is enriched primarily in GM12878 regions bound by Irf4. The known PU.1 motif matches Irf_disc5, and reciprocally PU.1_disc2 matches the Irf motif, consistent with the importance of PU.1 in hematopoietic development (Scott et al., 1994).

Beyond the discovered motif for Irf, several other discovered motifs (AP-1_disc2, CEBP_disc2, E2F_disc2, Pbx3_disc1, RFX5_disc2, and SP1_disc1-2) match the known NF-Y specificity (CCAAT). These discovered motifs are consistent with several known interactions of NF-Y. RFX5 promotes the cooperative binding between RFX and NF-Y (Villard et al., 2000), CEBPB and NF-Y interact in at least one promoter (Yu et al., 1995) and SP1 and NF-Y are known to interact (Roder et al., 1999). E2F_disc2 has particularly high enrichment in E2F4 datasets, consistent with the cooperative role E2F4 and NF-Y play in cell cycle regulation (Caretti et al., 2003).

STAT factors are involved in regulating number of growth related functions. We analyze STAT1, STAT2, and STAT3 here in the context of GM12878, HeLa-S3, MCF10A-Er-Src, and K562 cells. We find relatively consistent enrichment of the STAT full site (TTCCNGGAA), which STAT_disc1 matches, while finding weak enrichment for just the half-site (TTCC). We also find motifs involved in other proliferative functions including STAT_disc2, which is particularly enriched in STAT3 datasets and matches the TRE, consistent with STAT3 being one of the many interaction partners for AP-1 (Ivanov et al., 2001). STAT_disc3 matches the Irf consensus and has enrichment that is particularly high in STAT1 and STAT2 datasets stimulated by IFNa, highlighting the cooperativity of STAT factors and Irf in immune functions. STAT_disc4 is a match to the CEBPB motif and is found enriched in STAT3 datasets, consistent with the known cooperative role for these two factors (Choi et al., 2007).

Transcription factors with Ets domains are highly conserved and involved in several cellular processes (reviewed in Sementchenko and Watson, 2000). A number of TFs have discovered motifs that match the Ets consensus, including Egr-1_disc2, GATA_disc3, Nrf1_disc2, Pax-5_disc4, TR4_disc1, and Mef2_disc2. These discovered motifs are supported by known interactions between GATA and Ets in sea squirts (Rothbächer et al., 2007), Mef2 and the Ets factor PEA3 (Taylor et al., 1997), and TR4 with the Ets factor ELK4 (O'Geen et al., 2010). Moreover, Pax-5 and Ets factors have shared roles in the development of B-cells (Adams et al., 1992; Fitzsimmons et al., 1996). Looking at the discovered Ets motifs, we find that Ets_disc8 matches the known motif for Myb and the two have been known to cooperate, a relationship that is important in the context of certain cancers (Dudek et al., 1992).

THAP1 has two discovered motifs, both of which match the known YY1 motif (the first with additional specificity added by an apparent HNF4 motif). To our knowledge, the relationship between THAP1 and YY1 has not been directly observed, however, THAP1 has been known to associate with the coactivator HCF-1 (Mazars et al., 2010) and YY1 and HCF-1 are known to interact (Yu et al., 2010). Our result suggests that THAP1 and YY1, possibly with the addition of HNF4, may interact at least in the K562 cell line for which we have THAP1 binding data. Rad21_disc3 also matches YY1, suggesting an additional interaction.

Nanog, an important pluripotency transcription factor, has a known motif that is only weakly enriched (1.3-fold) in the bound regions and not discovered by our pipeline. We see much stronger enrichment for the known Pou5f1 and Pou2f2 motifs, for which we also find similar motifs (Nanog_disc2 and Nanog_disc4, respectively), consistent with their shared roles in pluripotency (Looijenga et al., 2003; Loh et al., 2006). The interaction of these factors is further supported by Pou5f1_disc2 matching the known Pou2f2 motif. Additionally, Nanog_disc2 and disc3 match the known motifs for TCF4 and TCF12, respectively, again consistent with the important role TCF proteins play in stem cells (Yi and Merrill, 2007).

CTCF plays a variety of vital roles in the organization of chromatin architecture (Phillips and Corces, 2009) and the motifs we discover matching the known CTCF specificity (Rad21_disc1, SMC3_disc1,2-4, CTCFL_disc1,10, ZBTB7A_disc1,2, SP2_disc3 and RXRA_disc2,5; some weakly) are largely compatible with this role. Rad21 is a highly conserved protein involved in DNA double strand repair (McKay et al., 1996) known to co-localize with CTCF (Wendt et al., 2008). Cohesion, of which SMC3 is a subunit, is brought to the chromatin by CTCF (Rubio et al., 2008). Further, while the function of the CTCF paralog CTCFL is not completely known, it does appear to be involved in imprinting through interaction with a histone methyltransferase (Jelinic et al., 2006).

A few of the discovered motifs contain additional specificity or appear to be combinations of multiple, distinct motifs. For example, Egr-1_disc4 appears to be a combination of multiple motifs (Egr-1, Ik-1, and a homeobox motif) and SETDB1_disc1 contains the Znf143 core sequence with significant additional specificity. The appearance of these motifs suggest highly specific "grammars" for these motifs that may require specific spacing and orientation of binding sites for functionality.

We find several additional enrichments of potential interest. Pbx3_disc2 matches the known Mrg motif, consistent with the known cooperative binding of Mrg and Pbx (Bischof et al., 1998). TAL1_disc1 matches GATA, with the potential connection that GATA and TAL1 are known to be important in hematopoiesis and vascular development (Kappel et al., 2000; Mouthon et al., 1993). Hsf_disc1 matches the known CEBP motif and has much higher enrichment in Hsf datasets (31-fold) compared to the known motifs for Hsf (<9-fold). Additionally, Egr-1_disc5, HNF4_disc5, Nrf1_disc3, Pax-5_disc2, RXRA_disc4/Pax-5_disc3, and SREBP_disc1 match the known motifs for Zic, Sox, SP1, Pax-2/3, Irf, and RFX5, respectively, suggesting additional previously uncharacterized interactions. Lastly, we find some motifs that show more ambiguous matches: BAF155_disc2 shows weak similarity to homeobox TGTAGT motif, TR4_disc2-3 weakly matches the known HNF4 motif, and Egr-1_disc3/SETDB1_disc2 matches the repetitive Nrf1 motif.

3.4.3 Key regulators revealed by cell line specific enrichments

Factors directly responsible for the establishment of enhancers, chromatin restructuring, or polymerase recruitment frequently exhibit binding that is highly cell line specific. Because most of these factors do not have their own sequence specificity, their binding is often correlated with that of regulators important for the specific cell line. We analyze several such factors (BCL, BDP1, CCNT2, Foxa, HDAC2, HMGN3, KAP1, p300, TATA, and TCF12) and find that key cell line regulators can be predicted by examining enrichments in cell lines specific datasets.

As a transcriptional coactivator, p300 interacts numerous TFs (reviewed in Chan and La Thangue, 2001) and it has been shown to have binding that can predict tissue-specific enhancers (Visel et al., 2009). Conversely, Foxa has a DNA binding domain and plays an important role in liver development and function (Costa et al., 2003) and is a pioneer factor responsible for priming chromatin for the binding of other factors (reviewed in Zaret and Carroll, 2011). Other proteins involved in chromatin restructuring include HDAC2, which transcriptionally represses through histone deacetylation (Johnson and Turner, 1999) and HMGN3 (Furusawa and Cherukuri, 2010). Further, two factor groups are directly involved in transcription including three RNA Pol3 subunits (BDP1, RPC155, and TFIIIC-110 and CCNT2, which is involved in the elongation of Pol2 (Peng et al., 1998).

Eight of these ten factor groups have at least one dataset in K562 (erythroleukaemia cells), and for four of these we discover motifs that match the GATA consensus which is then enriched specifically in the K562 datasets (BCL_disc5, CCNT2_disc1, HDAC2_disc1, and HMGN3_disc2). GATA has a known important role in K562 (Partington and Patient, 1999) and we also have previously found an association with GATA motifs and chromatin state derived enhancers for K562 cells (Ernst et al., 2011). We also find three additional motifs that have enrichment specific to the factor group's K562 dataset: BDP1_disc1, a 23-nt motif that contains the STAT consensus, HMGN3_disc1 which matches the TRE, and KAP1_disc2 which matches no known motif.

Likewise, for GM12878, an EBV mediated lymphoblastoid cell line, we find 3 discovered motifs that match the known Irf motif (BCL_disc4, p300_disc5, and TCF12_disc4). Irf-4 have been shown to be important in the establishment of these cell lines (Xu et al., 2008) and the family is an important player in immune cells (Paun and Pitha, 2007). This enrichment is also consistent with our previous study using epigenetic marks (Ernst et al., 2011), where we found Irf to be the strongest enriched motif in GM12878 specific enhancers. We also find GM12878 specific enrichment for motifs matching NF-kappaB (BCL_disc6) and Pou2f2 (TATA_disc9), consistent with the known biology of these factors (Corcoran et al., 1993; Baeuerle and Henkel, 1994).

Interestingly, we find that the TRE motif is found and enriched in a cell line specific manner for several factors, but for different cell lines. For example, HMGN3_disc1 is enriched in K562, BCL_disc2 has the highest enrichment in GM12878, KAP1_disc1 is only enriched in the HEK2932 and U2OS cell lines, and p300_disc7 has enrichment in the neuroblastoma cell line SK-N-SH-RA and HeLa-S3. This suggests that perhaps AP-1 or other factors recognizing TRE are selectively interacting with these proteins depending on the cell line.

The motifs we find specifically enriched in HepG2 (liver carcinoma) datasets match the known motifs for: Foxa (HDAC2_disc2, p300_disc3, and TCF12_disc2), HNF4 (Foxa_disc5 and HDAC2_disc5), and CEBP (p300_disc2,6), three key liver regulators (Costa et al., 2003; Lee et al., 2005). We find motifs with enrichments specific to H1-hESC which include matches to the pluripotency factor Pou2f2 (TATA_disc9), the near universally expressed repressor NRSF (BCL_disc3 and HDAC2_disc4), and key metabolic regulator Nrf1 (HDAC2_disc4). We find additional cell line specific enrichments for Foxa_disc3 (TCF12) in ECC-1, Foxa_disc4 (STAT) in both T-47D and ECC-1, and p300_disc2,6 (CEBP) and p300_disc4 (Ets) with enrichment in the HeLa-S3 dataset.

Even for these factors we find motifs that are consistently enriched across assayed cell lines for a given factor. Foxa_disc1, for example, matches the known Foxa motif, indicating that the coordinated binding Foxa participates in occurs in concert with Foxa's motif recognition. Most of the motifs we discover for RNA Pol2 machinery (TAF1, GTF2B, GTF2F1, and TBP) are enriched in all cell lines, including the known TATAAA motif (TATA_known4). Also, TATA_disc1, disc6, and disc 8 have consistent enrichment and match the known motifs for YY1 (which is known to be important in establishing transcription; Seto et al., 1991), NF-Y, and Ets. The top discovered motif BCL_disc1 matches the known Ets motif and is also enriched across datasets.

3.4.4 Novel motifs raise possibility of unknown regulators

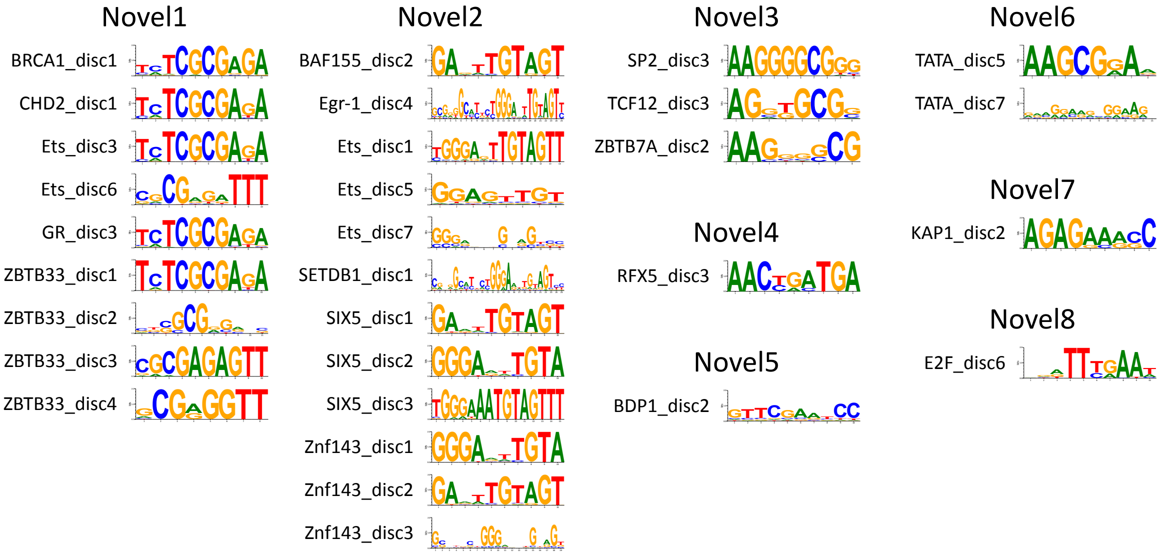


Figure 3-5: The eight putative novel motifs.

While we are able to putatively explain the majority of motifs we discover as either matches to previously known motifs or low complexity sequences, we do discover 30 putative novel motifs (Figure 3-5). We were able to place these into 8 groups on the basis of their similarity: Novel1 (BRCA1_disc1, CHD2_disc1, Ets_disc3,6, GR_disc3, and ZBTB33_disc1-4), Novel2 (BAF155_disc2, Egr-1_disc4, Ets_disc1,5,7, SETDB1_disc1, SIX5_disc1-3, and Znf143_disc1-3), Novel3 (SP2_disc3, TCF12_disc3, and ZBTB7A_disc2), Novel4 (RFX5_disc3), Novel5 (BDP1_disc2), Novel6 (TATA_disc5,7), Novel7 (KAP1_disc2), and Novel8 (E2F_disc6). These novel motifs were placed into a clusters first using the correlation criteria with 0.75 cutoff, but then manually annotated to include additional similar motifs particularly those found for the same factors.

Novel1 (using ZBTB33_disc1) is very highly enriched in at least one dataset for each of the factor groups for which it is found (BRCA1, CHD2, Ets, GR, and ZBTB33). All five factor groups except CHD2 have at least one known motif, and for each of these datasets Novel1 is more enriched in at least one dataset than any known motif (the result for GR is questionable because only one dataset has enrichment and that dataset has been independently flagged as problematic; see Kundaje et al., in review). The shared role of BRCA1 and CHD2 in DNA damage repair (Nagarajan et al., 2009; Deng, 2003) suggests that Novel1 may be involved in this or other shared roles for these factors and highlights the utility in shared motif enrichment even outside of motifs directly tied to a factor.

Similarly, for SIX5 we see only weak enrichment of the known SIX5 motif, and fail to discover a motif similar to it. However, Novel2 (using SIX5_disc1) shows over 100-fold enrichment for all three datasets (K562, GM12878, and H1-hESC). Novel2 also shows very high enrichment in datasets for which it was not found, including ATF3 (all datasets have >20-fold enrichment with GM12878 having 106-fold) and Nrf1 (all datasets have >30-fold enrichment). Moreover, the known Znf143 motif, which is 4-fold enriched in the one Znf143 dataset, is also not recovered, but Novel2 is 24-fold enriched. The breath of datasets sharing this motif suggests it may be recognized by an important regulator.

Like the known ZBTB7A motif, Novel3 (using SP2_disc3) is largely poly-G, which causes us to underestimate its enrichment due to our shuffling process. Despite this, however, it does show enrichment in several datasets, including for the factor groups for which it was discovered. Likewise, Novel4 (RFX5_disc3) shows moderate but consistent (2- to 6-fold) enrichment across the RFX5 datasets. The consensus is comprised of two of the same components as the known motifs (AAC and TGA), but ordered differently. Consequently, it may represent the binding specificity of, for example, an alternative isoform of RFX5. The remaining motifs (Novel5-8), were found for factors that show cell line specific enrichments. Consequently, these may represent specificities for regulators that are previously unidentified.

3.5 Conclusions

In this chapter we provide a systematic and comprehensive collection of motifs for hundreds of human TF binding datasets. TF binding can be complex, with a factor recognizing several or motifs, or binding in the apparent absence of any motif (reviewed in Farnham, 2009).

We also perform a thorough analysis of the results and make a number of biological contributions. We make specific predictions for interacting factors, validating several of them through a literature review. The several remaining interactions are appropriate for further follow-up. We also predict a number of novel motifs that are appropriate for further validation, either through computational means (e.g., conservation analysis) or experimentally.

This motif resource is used in several ENCODE papers, demonstrating its value for high throughput analyses. Our motifs are being matched at low stringency to find peaks that are void of any motif in order to understand the mechanism through which motif-less peaks are generated (Consortium, in review). The collection of known motifs and enrichment techniques we present here are also being used as a secondary validation of peaks (Landt et al., in review). Because having the motifs allows for more precisely determining the bases responsible for binding, these motifs enable analyzes involving population data (Spivakov et al., in review). Beal et al. (in review) are analyzing the conversation of motif instances produced by this dataset amongst the mammals. Kundaje et al. (in review) delve deeper into the relationship between pairs of motifs and use our motifs to predict factors that co-associate. Two other ENCODE papers also perform motif discovery: Wang et al. (in review) produce a non-redundant list of discovered motifs but do not perform an extensive analysis of the relationships between factors and Neph et al. (in review) use DNase footprinting data to predict relevant motifs.

Having a motif catalog is also the first step in predicting high quality computational targets of factors, which may allow the prediction of binding sites that were, for example, not found in the conditions assayed. Two popular strategies are used for this purpose. One is using the clustering of of motif instances for factors known to cooperate to form cis-regulatory modules (Berman et al., 2002; Schroeder et al., 2004). This resource is well-suited for this purpose because it naturally provides sets of motifs that are likely to cooperate. Instances can also be predicted using the strategy we discuss in the previous chapter, using many, closely related genomes.

Chapter 4

Computational discovery and characterization of microRNAs

This chapter presents novel machine learning algorithms for two related microRNA (miRNA) problems. First, we use evolutionary and structural features to build a *de novo* catalog of miRNAs. Second, we use additional features, including indirect ones, to predict the regulatory bases of the miRNA, which are essential for understanding the its target spectra. After presenting these algorithms we evaluate them on 12 *Drosophila* genomes and compare to experimental predictions and find unmatched performance. This work was done in collaboration with Alexander Stark and Leopold Parts, where I was responsible for the machine learning and several other computational aspects, but not the identification of putative hairpins and the generation of the features for each. This study was previously published in Stark et al. (2007a).

4.1 Introduction and historical context

MicroRNAs were first discovered in 1993 and have since become recognized as being an important class of regulators in animal genomes (reviewed in Bartel, 2004). They are endogenous, $\sim$22-nt RNAs that deactivate mRNAs through interacting with $3'$ UTRs and more limitedly, the coding region (Stark et al., 2007b; Bartel,

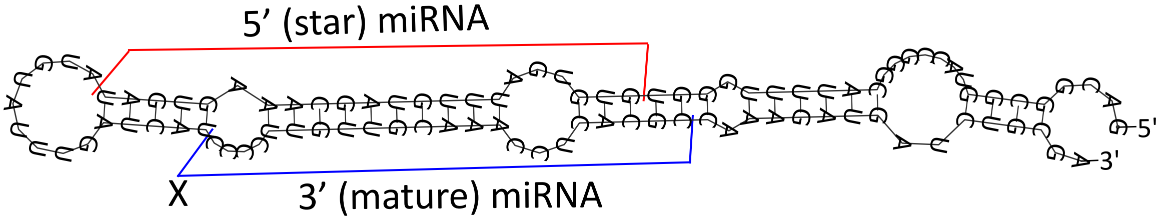


Figure 4-1: Minimum free energy (MFE) structure of dme-mir-988 as produced by RNAfold (Gruber et al., 2008). The mature and star sequences are indicated. The most highly sequenced RNA from the less sequenced arm (Figure 4-2) is referred to as the “star” sequence. The use of the “star” terminology has become somewhat less common recently in light of the high recorded frequency of both strands having functional mature sequences and the designation of the mature and star depending on the specific sample. Consequently, functional mature sequences have been named by their respective arm (i.e., either -5p or -3p) (Griffiths-Jones et al., 2008). X indicates the 5′ end of the mature miRNA which is the primary determinant of the binding of the miRNA.

2009; Schnall-Levin et al., 2010). In fly and worm genomes they are estimated to be about 1-2% as abundant as protein coding genes and each miRNA can target up to hundreds of genes (Bartel, 2009).

MicroRNAs are transcribed from potentially large genes (>> 1kb) to produce the primary or pri-miRNA (Figure 1-3). These are then processed twice: once in the nucleus by the protein Drosha to produce a hairpin structure (Figure 4-1; ∼85-nt) and then again by Dicer in the cytoplasm to produce the mature miRNA (∼22-nt) and the complementary star sequence. The star sequence is then generally discarded while the mature is incorporated into the RNA-induced silencing complex (RISC). Despite their importance in understanding the regulation of miRNAs, annotation of pri-miRNAs has not been amenable to purely computational approaches and is only now becoming experimentally possible through analysis of chromatin modifications (Corcoran et al., 2009).

On the other hand, determination of the miRNA hairpins has been an active area of research both through experimental and computational approaches (reviewed in Berezikov et al., 2006). The first miRNAs were found using forward genetics. Later, experimental protocols were developed to sequence size selected RNAs to identify the products of miRNA processing. These are then matched to the genome and miRNA hairpins are identified by the examination of the processed RNA products (Figure 4-2). Putative miRNAs (e.g., predicted from computational methods), could also be confirmed through Northern Blot analysis (Lai et al., 2003), although with a relatively high false positive rate (this technique is a source of the non-sequenced miRNAs in miRBase).

Because the number of miRNAs is relatively small (e.g., ∼250 for fly, ∼1500 for human; Kozomara and Griffiths-Jones, 2010) manual annotation of sequenced reads is possible, and was performed, for example, in Ruby et al. (2007). More recently, miRDeep (Friedlander et al., 2008) has permitted the automated processing of small RNA sequence datasets for identifying miRNAs. We also developed a partially automated pipeline for identifying miRNAs from small RNA sequencing reads and successfully applied it to find platypus (Murchison et al., 2008) and later Tasmanian devil (Murchison et al., 2010) miRNAs, the later without an available devil genome.

In contrast to these experimentally driven approaches, in this chapter we present a *de novo* computational approach for the prediction of miRNAs, a problem that has been previously investigated (Lai et al., 2003; Bentwich et al., 2005; Lim et al., 2003a,b). These require comparative information because, as we will see, miRNA-like hairpins are highly abundant in the genome and structural features alone are insufficient for their identification. In recent years, advances in technology have led to sequencing largely supplanting computational methods for miRNA prediction. Consequently, this study and a parallel one (Ruby et al., 2007) continue to be the most recent large scale efforts to computationally predict miRNAs. Moreover, this study continues to be relevant even today because: (1) 17 of the predicted miRNAs have not yet been sequenced and may represent real miRNAs that have low expression or are found only in cell types or conditions not experimentally assayed, and (2) as we will see the features that are ultimately used are informative to miRNA biology.

```
GGGCAUUUUGGGUGUGUGAUUUGUAGCAAAGUGAUAUGUAUUUGAUCAUCCCCUUGUUGCAAACCUCACGCCAAAGAUGAUCU
+++++++++++++                                                                      CGGUACCGGGCAUUUUGGGU           4
             +++++++++++++++++++++++                                               GUGUGAUUUGUAGCAAAGUGAUA        3
             ++++++++++++++++++++++                                                GUGUGAUUUGUAGCAAAGUGAU         10
             +++++++++++++++++++++                                                 GUGUGAUUUGUAGCAAAGUGA          10
             ++++++++++++++++++++                                                  GUGUGAUUUGUAGCAAAGUG           6
              +++++++++++++++++++++                                                UGUGAUUUGUAGCAAAGUGAU          2
              ++++++++++++++++++++                                                 UGUGAUUUGUAGCAAAGUGA           1
                                                +++++++++++++++++++++++            UCCCCUUGUUGCAAACCUCACGC        4
                                                ++++++++++++++++++++++             UCCCCUUGUUGCAAACCUCACG         16
                                                 +++++++++++++++++++++++           CCCCUUGUUGCAAACCUCACGCC        10
                                                 ++++++++++++++++++++++            CCCCUUGUUGCAAACCUCACGC         5747
                                                 +++++++++++++++++++++             CCCCUUGUUGCAAACCUCACG          5617
                                                 ++++++++++++++++++++              CCCCUUGUUGCAAACCUCAC           510
                                                 +++++++++++++++++++               CCCCUUGUUGCAAACCUCA            173
                                                 ++++++++++++++++++                CCCCUUGUUGCAAACCUC             2
                                                  +++++++++++++++++++++            CCCUUGUUGCAAACCUCACGC          12
                                                  ++++++++++++++++++++             CCCUUGUUGCAAACCUCACG           1
                                                   +++++++++++++++++++++           CCUUGUUGCAAACCUCACGCC          5
                                                   ++++++++++++++++++++            CCUUGUUGCAAACCUCACGC           15
                                                   +++++++++++++++++++             CCUUGUUGCAAACCUCACG            2
                                                    +++++++++++++++++++            CUUGUUGCAAACCUCACGC            6
                                                    ++++++++++++++++++             CUUGUUGCAAACCUCACG             1
                                                     ++++++++++++++++++            UUGUUGCAAACCUCACGC             2
```

Figure 4-2: Sequenced reads matched to genomic hairpin of dme-mir-988. The number of reads for each sequence is indicated. The characteristic read abundance pattern for miRNAs is seen: reads match the hairpin in two groups with the mature (here 3′) having much higher abundance than the star sequence. Other notable features are the higher variability in the 3′ end of the mature sequences and occasional reads from the extreme ends of the miRNA; reads are occasionally also recovered from the loop but are absent here.

4.2 MicroRNA hairpin prediction

4.2.1 Random forests for miRNA gene finding

We developed a machine learning framework for the purposes of predicting miRNA hairpins in the *Drosophila melanogaster* genome. To predict miRNA-like hairpins, we first ran RNAfold from the Vienna package (Hofacker et al., 1994) on 120-nt windows (overlap of 90-nts) and trimmed each window to the end of any hairpins. The resulting fold was used to infer the arms and loop of each hairpin. As a lenient prescreening, we removed all hairpins shorter than 63-nts, with an arm of less than 20-nts or with less than 70% arm base-pairing. We were left 760,355 potentially overlapping list of putative miRNA hairpins. Amongst these 760,000 hairpins we found all 60 sequenced *Drosophila melanogaster* miRNAs from miRBase release 9.0 (Griffiths-Jones et al., 2008), which we use as our positive set. Because the number of estimated fly miRNAs is only in the hundreds (Lai et al., 2003), we expect the vast majority of the 760,000 hairpins to be spurious. Con-

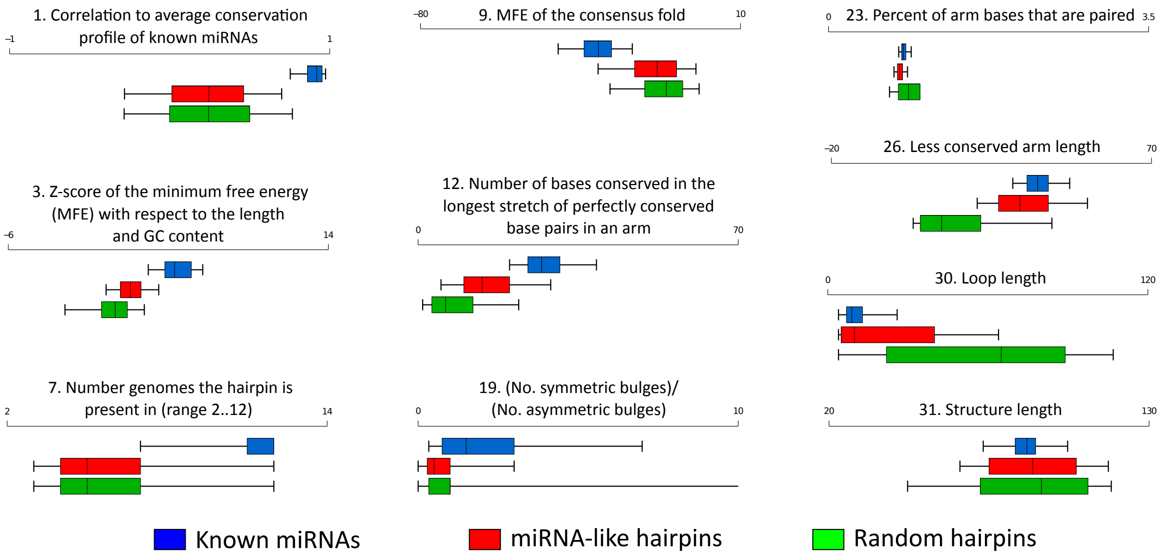


Figure 4-3: Selected hairpin features from Table 4-1. Boxes show 25th percentiles, whiskers show 95th percentiles. We see that known miRNAs have several structural features that partially distinguish them including a lower (i.e. more significant) MFE and a higher number of symmetric vs. asymmetric bulges. However, the conservation-based features are those that are particularly discriminative including the correlation to the conservation profile for known miRNAs and the number of genomes the hairpin is found in.

sequently, while a true negative set cannot be identified, we expect that the bulk properties of all the hairpins to be dominated by those of non-miRNA hairpins.

For each melanogaster hairpin sequence, we selected the best BLAST (Altschul et al., 1990) match with E-value $\leq 1 \times 10^{-5}$ in each of the 11 other genomes (CAF1 assemblies). We then added 50-nt flanking sequenced and produced a multiple alignment using ClustalW (Thompson et al., 1994). For historical reasons we did not use whole genome alignments for this purpose (high quality ones were not yet available). However, we suspect this procedure would work roughly as well (or perhaps even better) if run on, for example, the UCSC genome browser alignments because they cover the known miRNAs well and may be more accurate due to their use of synteny and intermediate species.

Each hairpin was then scored using various structural and conservation-based properties (Table 4-1 and Figures 4-3 and 4-6). We see that the best feature (Figure 4-4) only provides 327-fold enrichment for known miRNAs at the cutoff with

Feature	Known miRNAs	miRNA-like hairpins	Random hairpins	Enrichment
1. Correlation to average conservation profile of known miRNAs	0.75 - 0.97	-0.28 - 0.7	-0.28 - 0.77	327.5
2. Match fraction for the less conserved arm	5.9 - 12	1.8 - 8.9	1.6 - 8.1	42.3
3. Z-score of the minimum free energy (MFE) with respect to the length and GC content	2.7 - 6.2	0.11 - 3.4	-2.5 - 2.5	39.2
4. Match fraction for the more conserved arm	6.1 - 12	2.5 - 10	2.5 - 10	16.1
5. Mismatch fraction for the region flanking the more conserved arm	1.9 - 6.4	0.05 - 3.7	0.05 - 3.7	13.4
6. Mismatch fraction for the loop	0.33 - 5.5	0 - 2.2	0 - 2.1	10.9
7. Number genomes the hairpin is present in (range 2..12)	7 - 12	3 - 12	3 - 12	9.6
8. Mismatch fraction for the region flanking the less conserved arm	2.4 - 6.2	0.1 - 4.6	0.09 - 4.7	6.6
9. MFE of the consensus fold	-41 - -20	-30 - -2.5	-27 - -1.6	6.3
10. Average difference between the MFE of the consensus fold and the individual fold	0.59 - 9.4	1.9 - 22	1.3 - 21	6.2
11. (less conserved arm conservation)/(loop conservation)	0.99 - 1.7	0.63 - 1.2	0.53 - 1.2	5.2
12. Number of bases conserved in the longest stretch of perfectly conserved base pairs in an arm	20 - 39	5 - 29	1 - 22	5.0
13. Match fraction for the region flanking the less conserved arm	3.9 - 8.6	1.7 - 7.9	1.7 - 7.6	4.7
14. (more conserved arm conservation)/(loop conservation)	1 - 1.8	0.86 - 1.4	0.87 - 1.3	4.1
15. Match fraction for the loop	4.1 - 12	2 - 10	2.2 - 9.7	3.9
16. (MFE of consensus fold/average MFE of the individual ortholog folds)	0.68 - 1	0.15 - 0.96	0.12 - 0.97	3.7
17. Match fraction for the region flanking the more conserved arm	3.9 - 8.7	2 - 8.8	1.9 - 8.6	3.4
18. Mismatch fraction in less conserved arm	0.036 - 1.8	0.19 - 3.4	0 - 4.4	3.3
19. (No. symmetric bulges)/(No. asymmetric bulges)	0.33 - 7	0.000025 - 3	0.00005 - 10,000	3.0
20. Mismatch fraction in more conserved arm	0 - 1.3	0 - 1.4	0 - 1.8	2.6
21. Number of paired bases in the best stretch of 22 base pairs	18 - 22	16 - 21	3 - 19	2.6
22. Indel fraction for the loop	0 - 0.16	0 - 0.17	0 - 0.14	2.4
23. Percent of arm bases that are paired	0.77 - 0.9	0.72 - 0.86	0.67 - 1	2.3
24. Number of bases that need to be removed to make all internal bulges symmetric	1 - 7	1 - 13	0 - 10	2.0
25. Indel fraction for the region flanking more conserved arm	0.012 - 0.17	0 - 0.14	0 - 0.15	2.0
26. Less conserved arm length	31 - 47	21 - 52	3 - 42	1.7
27. More conserved arm length	31 - 47	21 - 52	3 - 42	1.7
28. Indel fraction for the less conserved arm	0 - 0.036	0 - 0.23	0 - 0.26	1.7
29. Indel fraction for the region flanking less conserved arm	0.018 - 0.17	0 - 0.2	0 - 0.2	1.5
30. Loop length	4 - 26	4 - 64	4 - 110	1.4
31. Structure length	73 - 100	65 - 120	47 - 120	1.4
32. Number of substructures	0 - 0	0 - 1	0 - 2	1.4
33. Indel fraction for the more conserved arm	0 - 0.02	0 - 0.078	0 - 0.07	1.2
34. Number of internal bulges	3 - 8	2 - 9	0 - 6	1.2

Table 4-1: Features used for hairpin discovery. Ranges are 5-95% intervals; Hairpins indicate the 760,000 genomic miRNA-like hairpins and Random indicates random genomic hairpins (without the 'miRNA-like' filter). Enrichment indicates the enrichment of known hairpins when using the cutoff on the feature with the highest information content. All indicated features were used by random forest algorithm, however, these features come from a larger pool that were manually pruned for redundancy. We did not make substantive changes to the predictions after intersection with miRNAs from Ruby et al. (2007).

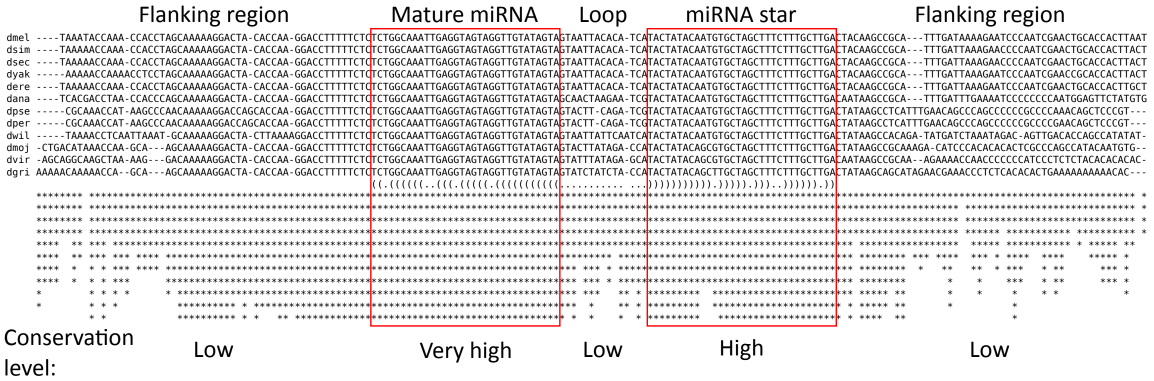


Figure 4-4: Conservation profile of dme-let-7, which roughly follows the average conservation profile of all sequenced miRNAs. Very high conservation is seen in the arm corresponding to the mature miRNA and high conservation is seen in the opposite arm. The loop and flanking regions tolerate mutations. Correlation to the average profile is the most distinguishing single feature.

highest information content — when recovering 42 known miRNAs it also predicts 1,625 additional miRNAs — which is significantly greater than the current prediction of the number of conserved miRNAs and, as we will see, more than we predicted when combining features.

In order to combine these features, we employed a machine learning algorithm similar to Random Forests (Breiman, 2001) but customized for our purposes. Using the combined conservation and structural feature set (Table 4-1 and Figure 4-6), 500 decision trees were trained on a positive training set of 60 miRNAs and a different randomly selected negative set of 250 of the remaining miRNA-like hairpins.

Decisions trees were trained using the standard C4.5 algorithm on continuous variables (Quinlan, 1996). We start by placing all the samples in one node. Nodes are then iteratively selected (the order is unimportant) and split according to the feature and cutoff combination that results in the in the highest information gain. This simplifies to maximizing $\sum_{i\in 1,2}\sum_{j\in 0,1} |x_{i,j}| \log \frac{|x_{i,j}|}{|x_i|}$, where $|x_{i,j}|$ indicates the number of samples in partition i that are positive ($j = 1$) or negative ($j = 0$). This is repeated until there is no additional information gain at any node for a simple split. While we experimented with pruning of the decision trees (merging nodes that were otherwise homogeneous), we found that it did not improve performance. The resulting 500 decisions trees "vote" and the aggregate score is the fraction which vote that a sample is a miRNA, where the vote for each tree is the majority value at the leaf the sample resolves to.

The intent of this procedure was to capture the variety of known miRNAs while not excluding potentially real genomic hairpins. The final score for a hairpin was derived through cross validation where, in order to avoid optimistic performance due to over-fitting, we placed all overlapping hairpins and all known miRNAs that are products of recent duplication into the same cross validation group. The 760,000 putative hairpins were also scored using cross validation by being randomly placed into one of two cross validation groups. From all hairpins that overlap on the same strand, only the hairpin with the highest score is kept. In some cases, this led to known miRNAs having a slightly revised hairpin selected.

Score	Locus	Name	Host gene	Species	Predicted mature	Targets
1.000	3R 27091338 27091410 -	Novel-8		12+	CAAATTAACTGCGACATGGC	479
0.986	3L 4989729 4989829 +	Novel-17		10+	AAAATTATGCGGAAACGGAAGC	519
0.980	3L 10936322 10936415 +	Novel-21		12+	TCGTCGCATGCGCGTGATCAAC	26
0.980	3R 23797295 23797385 -	Novel-22	betaTub97EF	9	TTTATTGCGGCCTGGCCTGACA	521
0.978	2R 10136644 10136747 +	Novel-23		12+	GAAAGAATAAGAACGGCCAACT	105
0.976	3R 22570390 22570454 +	Novel-24		12+	TCAATCAAATCACATGACTGCT	98
0.976	3R 24822601 24822689 -	Novel-25	CG1443	12+	TGCATTTAAGCCAATTAGCATA	292
0.974	3L 7188122 7188227 -	Novel-27		5	TGAGTCCTTTCACTGGCCACTC	23
0.968	2L 11749802 11749899 -	Novel-28		8+	TGCTTTGAGGTTTATTAGCTGC	109
0.964	3R 16281787 16281884 +	Novel-31		12+	AATGTCATTAAATTCTCATACA	68
0.962	X 15110474 15110559 +	Novel-32		11+	TTTTATTTGTGTCACTGAGTGG	1399
0.960	3R 21923504 21923596 +	Novel-33		12+	TTTGTTCGAGTTGACGTTTGGA	178
0.960	X 18015331 18015398 -	Novel-34		10	TACATAATGTCTCTGTAGGCC	320
0.958	2L 3858079 3858150 +	Novel-35		12+	AATTTAATGTGTCGGCGTGTTT	482
0.958	2R 13077473 13077566 +	Novel-36	Klp54D	8+	TGTTCTCTCCCATTTCTGACTC	59
0.952	2L 15654271 15654348 -	Novel-39		11+	TAATTGCCTGTAAACATAAAGG	146
0.952	3L 7528550 7528614 -	Novel-41		12+	TACTTTTACTTTCATTTATCAA	193

Table 4-2: Predicted, non-validated miRNAs (score ≥ 0.95). While these 17 putative miRNAs were not validated, they may represent miRNAs expressed in conditions or cell types not assayed by sequencing efforts. "Locus" indicates positions in dm2 assembly; "Species" indicates the number of species the hairpin is found in with "+" indicating the mature is fully conserved in all species. Number of targets from Kheradpour et al. (2007).

4.2.2 Recovery of known and newly sequenced miRNAs

Ultimately, the combination of these scores in this machine learning framework is able to enrich for known miRNAs at ~4,500-fold (Figure 4-5 and 4-6). After collapsing overlapping hairpins, 101 distinct predicted miRNAs had a score of

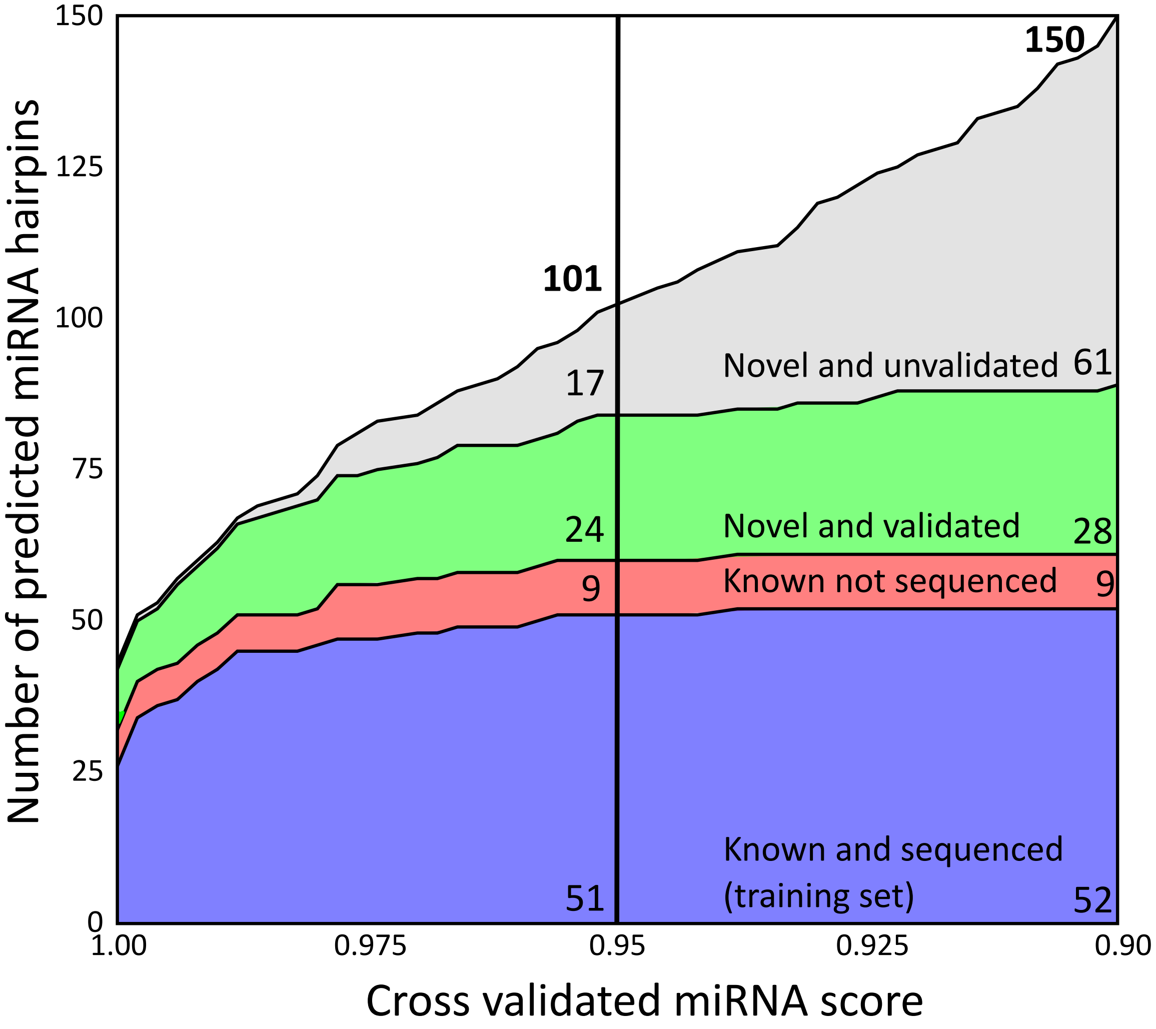


Figure 4-5: Identity of top scoring 150 hairpins. We find that at either the 0.95 or more lenient 0.90 cutoff, most of the top ranking hairpins are either previously sequenced or sequenced in a study parallel to this one (Ruby et al., 2007).

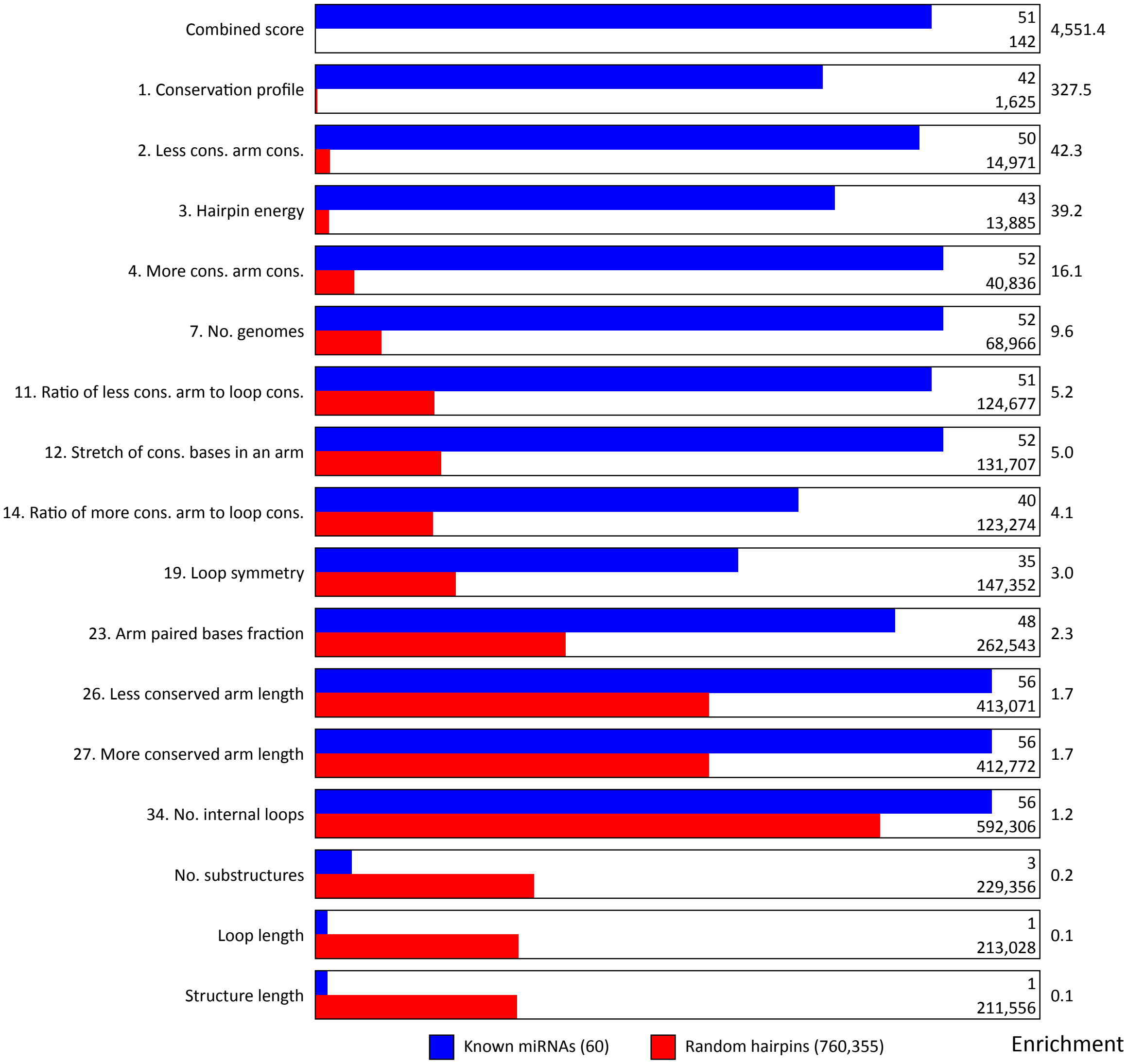


Figure 4-6: Comparison of machine learning score to hairpin features in terms of the number of previously known miRNAs and number of random hairpins recovered at the cutoff with the highest information. Combining the features leads to a dramatically higher enrichment for known miRNAs, sufficient for the *de novo* identification of miRNAs. Features corresponding to those used in machine learning indicated by number (see Table 4-1).

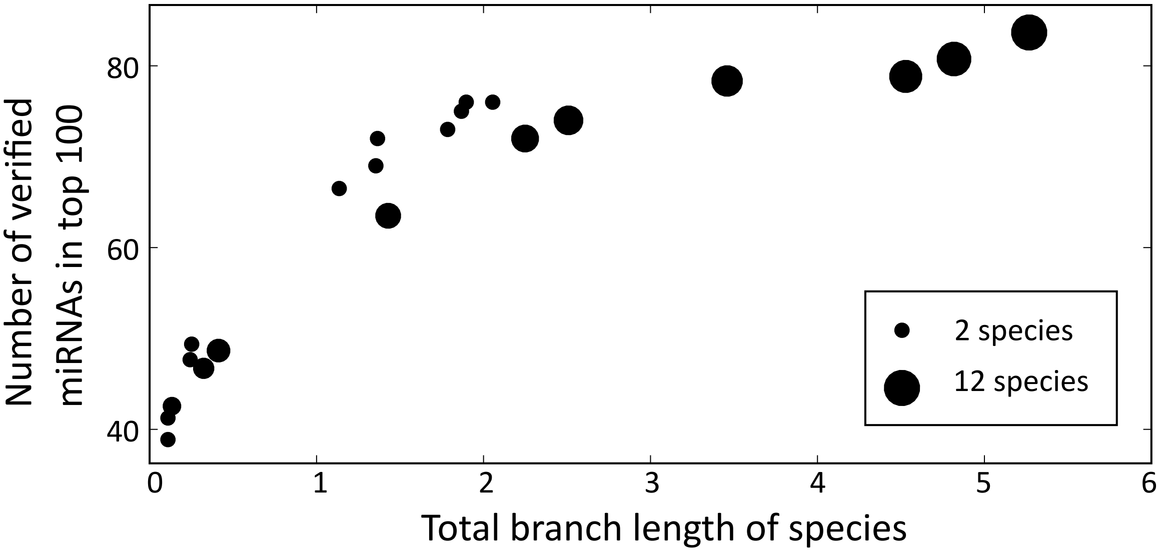


Figure 4-7: Performance of hairpin prediction algorithm on different subsets of species. Size of circle indicates number of species used (from 2 to 12).

at least 0.95, our more stringent cutoff. Among these we recover 51 of the 60 miRNAs used in our training set. Moreover, we find 9 of 18 miRNAs that were included in miRBase but not validated through sequencing (additional sequencing led to 4 of the 18 being discarded; Ruby et al., 2007). 24 of the remaining 41 predicted hairpins are validated through comparison with sequencing reads produced in parallel to this study (Ruby et al., 2007). The remaining 17 miRNAs predictions (Table 4-2) have diverse structural and evolutionary properties indicative of miRNA function and may be expressed at low levels or in conditions not assayed.

To investigate the scaling of performance with a variable number of *Drosophila* species, we reran our algorithm (learning new models) using different subsets of the 12 available species and looked at the number of verified miRNAs in the top 100 predictions (Figure 4-7). As in Chapter 2 with motif instance prediction, performance roughly scales linearly with branch length and does not appear to have saturated with the 12 fly genomes. We recover 84 miRNAs when using all 12 species, however, we are able to recover as many as 76 miRNAs using only 2 species (*D. melanogaster* and the most distant species, *D. grimshawii*). Also, using a single informant species tends to perform better than using several species with the same combined branch length, suggesting room for improvement in comparative features which may not be fully capitalizing on the phylogenetic relationship off the species.

These analysis also allow us to compare to Lai et al. (2003), that also made *D. melanogaster* miRNA predictions, but only using *D. pseudoobscura* as an informant. Lai et al. predicted a total of 208 miRNAs, which in retrospect contains 73 verified miRNAs. In contrast, our algorithm run using only *D. pseudoobscura* as an informant recovers 84 miRNAs in the top 208 predictions, a significant increase. This difference is attributable to one of three differences between our approaches: the specific features we use, the machine learning algorithm we employ, or the availability of additional training miRNAs.

Ruby et al. (2007) also made computational miRNA predictions contemporaneously to ours and found 75 validated miRNAs in their top 100 predictions compared to the 83-84 we found in our top 100. However, it is important to note that: (1) the 75 includes 31 miRNAs that were used in training, (2) Ruby et al. used only 6 species at a total branch length of 3.91. At a similar branch length (but different species set) we recover between ∼78-79 validated miRNAs. Moreover, our top 101 predictions contain 41 miRNAs not in miRBase, of which 24 (59%) are experimentally validated, while Ruby et al. predicts 45 miRNAs not in miRBase, of which only 20 are validated using the same experimental dataset.

4.3 Accurately predicting the mature miRNA

4.3.1 Relevant features of mature miRNAs

The functional specificity of a miRNA is almost completely determined by the first 8 bases of the mature sequence, particularly positions 2-8 (reviewed in Bartel, 2009). Consequently, while discovery of miRNA hairpins is important for understanding the evolution and genomic context of a specific miRNA, knowing the mature sequence is vital to determining their functional significance. Once the precise $5'$ cleavage position is known, the reverse complement of positions 2-8 can be treated as motifs to be matched against the $3'$ UTRs of genes (see Chapter 2). Highlighting the biological pressure to produce the appropriate mature $5'$ end, the $5'$ end processing of mature sequences is more accurate than that of either $3'$ ends or of the $5'$ end of the star sequence (Ruby et al., 2006).

While sequencing the mature miRNAs (Figure 4-2) is the gold standard for identifying the $5'$ end, several properties also distinguish them and have been used for their prediction (Table 4-3 and Figure 4-9), although with limited power (e.g., Lai et al., 2003). First, the mature sequence is almost always perfectly conserved in nearby species. Second, as has been previously observed (Lau et al., 2001), while $\sim$30% of miRNA bases are uridine (U), 78% of mature miRNAs start with a U. Further, we observe that the number of paired bases in a window near the mature start is constrained.

Mature miRNAs also exhibit features that while not observable by the processing machinery, never-the-less distinguish their $5'$ end (Figure 4-8). In particular, it has been observed that the sequences that recognize miRNAs are preferentially conserved in $3'$ UTRs of genes that are potential targets for miRNAs (Lewis et al., 2005; Stark et al., 2005; Xie et al., 2005). Moreover, these 7-mers are avoided in the $3'$ UTRs of ubiquitously expressed genes (Stark et al., 2005; Farh et al., 2005), presumably because otherwise they would be inadvertently down regulated in cell types where the corresponding miRNAs are expressed. Because 7-mers starting at adjacent bases share a significant fraction of their targets, the conservation and avoidance profile for adjacent 7-mers is highly correlated. We account for this when combining the scores, as we will see, to accurately predict the mature $5'$ cleavage.

4.3.2 Using an SVM to find the $5'$ cleavage site

We computed relevant features (Table 4-3) for each position in each miRNA hairpin, excluding those that would not permit sufficient space for a mature miRNA

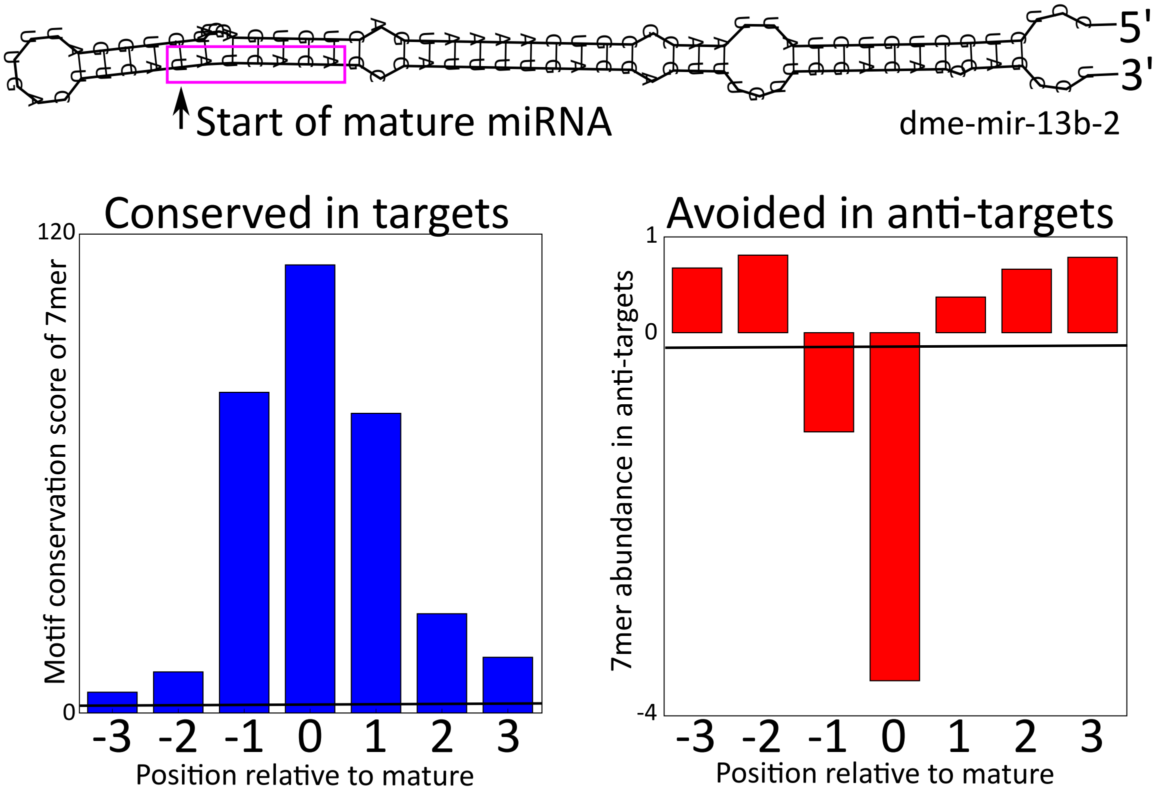


Figure 4-8: Indirect features used for the identification of the mature miRNA. The average profile across all miRNAs is similar to what is seen for this particular miRNA. MCS (Kellis et al., 2003, 2004; Xie et al., 2005) of the 7-mer matching the start of the mature miRNA is higher in 3′ UTRs than for an average 7-mer (indicated as the solid line). Conversely, these 7-mers are avoided in global anti-target genes (Stark et al., 2005).

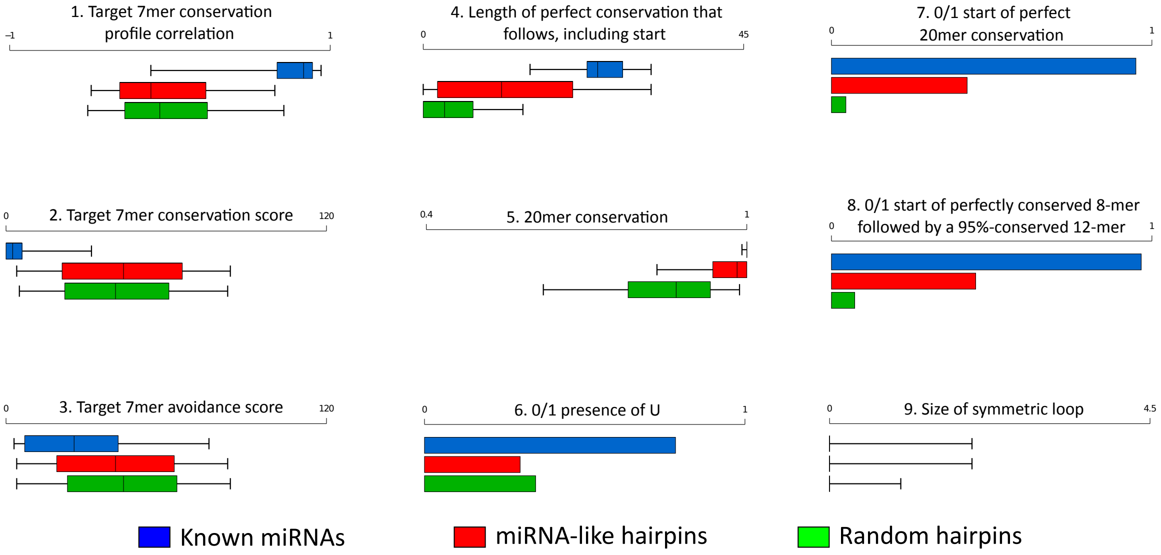


Figure 4-9: Selected mature features from Table 4-3. Boxes show 25th percentiles, whiskers show 95th percentiles.

Feature	Known $5'$	Other positions	Enrichment
1. Target 7-mer conservation profile correlation	-0.12 - 0.94	-0.49 - 0.65	15.2
2. Target 7-mer conservation score	5.9 - 110	-0.54 - 24	11.04
3. Target 7-mer avoidance score	-1.5 - 4.6	-3.5 - 3.7	4.11
4. Length of perfect conservation that follows, including start	15 - 32	0 - 32	3.54
5. 20-mer conservation	0.99 - 1	0.83 - 1	2.85
6. 0/1 presence of U	0.78	0.30	2.62
7. 0/1 start of perfect 20-mer conservation	0.95	0.42	2.24
8. 0/1 start of perfectly conserved 8-mer followed by a 95%-conserved 12-mer	0.97	0.45	2.15
9. Size of symmetric loop	0 - 2	0 - 2	1.9
10. Distance from terminal loop	0 - 9	-5 - 20	1.46
11. Distance from start of the hairpin	3 - 16	-2 - 23	1.43
12. Overlap length with the loop region	0 - 0	0 - 5	1.24
13. 0/1 loop at 1 position	0.27	0.22	1.21
14. Number of paired bases in window of 7	3 - 7	2 - 7	1.18
15. 0/1 loop at 2 position	0.08	0.21	1.16
16. 0/1 loop at -2 position	0.18	0.26	1.11
17. Size of overlapping bulged loop	0 - 0	0 - 6	1.08
18. 0/1 loop at -1 position	0.22	0.25	1.04
19. Number of paired bases in window of 3	1 - 3	0 - 3	1.03
20. Number of paired bases in window of 5	2 - 5	1 - 5	1.02
21. 0/1 loop at current position	0.23	0.23	1

Table 4-3: Features used for mature $5'$ prediction. Ranges are 5-95% intervals, except for binary statistics which are given as fractions.

(for positions in the left arm, we required at least 15-nts before the start of the loop and for positions in the right arm, we allowed no more than a 3-nt overlap with the loop and required at least 18-nts before the end of the hairpin). 7-mer conservation is assessed using motif-conservation scores (MCS) of 7-mers calculated on all annotated $3'$ UTRs, as previously described (Kellis et al., 2003, 2004; Xie et al., 2005) and the avoidance score is computed for each 7-mer using the $3'$ UTRs of global anti-target genes (Stark et al., 2005) by computing the deviation relative to all genes by Z-scores.

Within each hairpin, we linearly normalized each feature to be from 0 to 1 and marked each known mature site as a positive and all remaining permissible sites as negative. We augmented the features for each position with the features of the previous and next positions, permitting the machine learning algorithm to take advantage of, for example, 7-mer conservation scores for adjacent positions. We used SVMlight (Joachims, 1999) to train an SVM with default parameters (linear kernel and positive gain 1) on all the permissible locations from all the known hairpins and predicted the mature location by taking the permissible location in

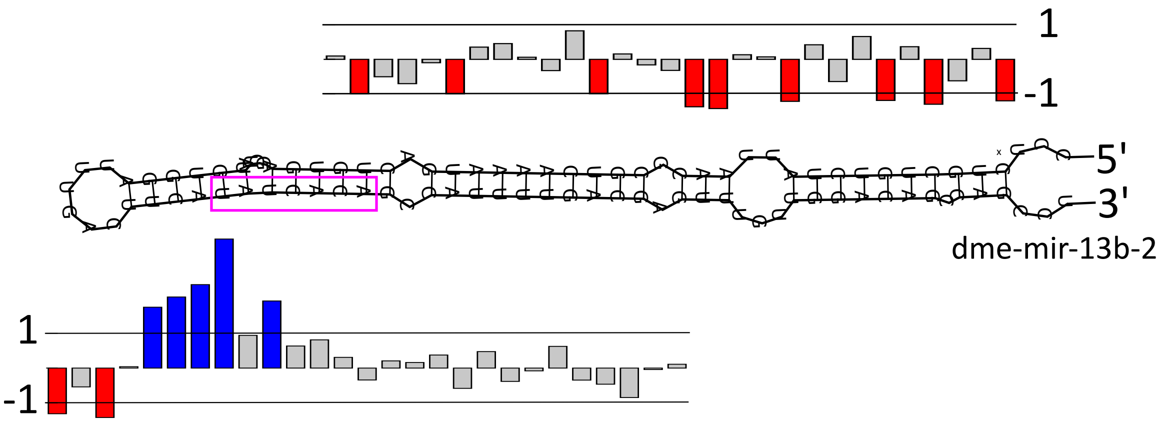


Figure 4-10: Example of a score profile for a miRNA where the known mature 5′ cleavage is predicted. Scores are shown for all permissible locations.

each hairpin with the highest SVM score. SVM scores are normalized to have mean 0 and standard deviation 1 within each miRNA. As with the hairpin prediction, we use cross validation to obtain a predicted mature location for each miRNA (each hairpin is only scored by models trained on known hairpins that exclude it and all its family members).

4.3.3 Performance of mature prediction

Using cross validation, we obtained the correct 5′ end for 47 out of 60 training miRNAs (78%) and were within 1-bp for an additional 4 (e.g., Figure 4-10). An additional 14 previously unsequenced miRNAs were in miRBase that now had sequencing evidence and consequently a reliably known 5′ end. We disagreed with the previous annotation in 9 cases. For 6 of those 9 we were more accurate than the previous annotation: for 4 we were exactly right, and for the other two within 1-nt. In 5 cases we agreed with the previous annotation, 4 of which were confirmed by sequencing and the other both the previous annotation and ours was off by 1.

4.4 Biological insights

Beyond the computational contributions of our machine learning approach, the specific useful features and the predicted miRNAs expand our knowledge of miRNA biology.

4.4.1 Multiple functional mature products from one locus

In the hairpin analysis above, we considered whether or not we were able to predict the correct *locus* (i.e., ignoring the strand). Initially it may seem to be difficult to distinguish between two strands of a miRNA, and indeed evolutionary features generally do not differ between the two strands. However, because RNA structure permits G-U base pairing, RNAs derived from opposite DNA strands generally have different folds.

Consequently, while 51 of the 60 training miRNAs had a score ≥ 0.95 on the positive strand, only 21 had a hairpin on the negative strand whose score was ≥ 95. Moreover, only 4 of the training miRNAs had a higher score on the sense strand compared to the anti-sense strand. Interestingly, 4 miRNAs (dme-iab-4, mir-307, -124, and -305) also had reads on the opposite strand, albeit at much lower levels than on the positive strand. For all four of these cases both strands had scores of at least 0.97. Experimental follow by us and others of one such case, dme-iab-4, led to the prediction of a functional anti-sense miRNA in the heavily studied Hox cluster (Tyler et al., 2008; Stark et al., 2008; Bender, 2008).

We also found evidence for functionality of alternative mature sequences. In particular, when we accurately predicted the mature $5'$ end, an average of 90% of reads intersecting a miRNA supported that start compared to 78% when we failed ($P = 6 \times 10^{-3}$). A striking example of this is dme-mir-964, where the position we predict (shifted by 1-nt) is supported by more than half as many reads that support the annotated start (Figure 4-11). Moreover, we found a higher number of reads for miRNAs that had high scoring star sequence: the 10 miRNAs with the highest star SVM score had 9% of their reads supporting the star, compared

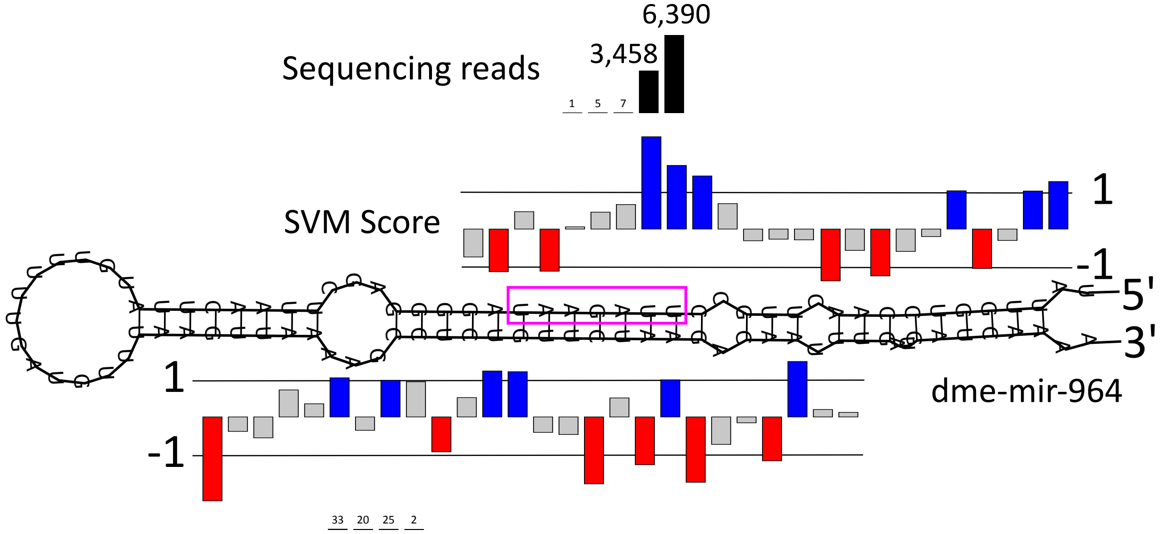


Figure 4-11: 5′ cleavage scores for dme-mir-964 show maximum over an alternative site which has 54% as many reads as the most abundant RNA from the mature arm.

to 2% for other miRNAs.

4.4.2 Novel hairpins give alternative explanation for transcripts

The validated dme-mir-996 overlaps with the 5′ UTR of CG31044 and the unvalidated Novel-60 overlaps with the coding region of CG33311. As part of an independent effort to update the coding exon catalog of *D. melanogaster* (Lin et al., 2007b), these two genes were determined to lack evolutionary properties associated with protein coding exons. In particular, alignments of their exons were littered with frame shifting indels, stop codons and non-synonymous mutations. Consequently, our analysis provides an alternative explanation for these previously identified transcripts as pri-miRNAs. Further, it highlights the benefits of performing an unbiased, *de novo* search, that unlike previous approaches did not explicitly exclude exons.

4.4.3 MicroRNA expression, conservation, and targets correlated

We expect the newly sequenced miRNAs (Ruby et al., 2007) to have a lower overall expression level due to the necessity to employ more sensitive methods for

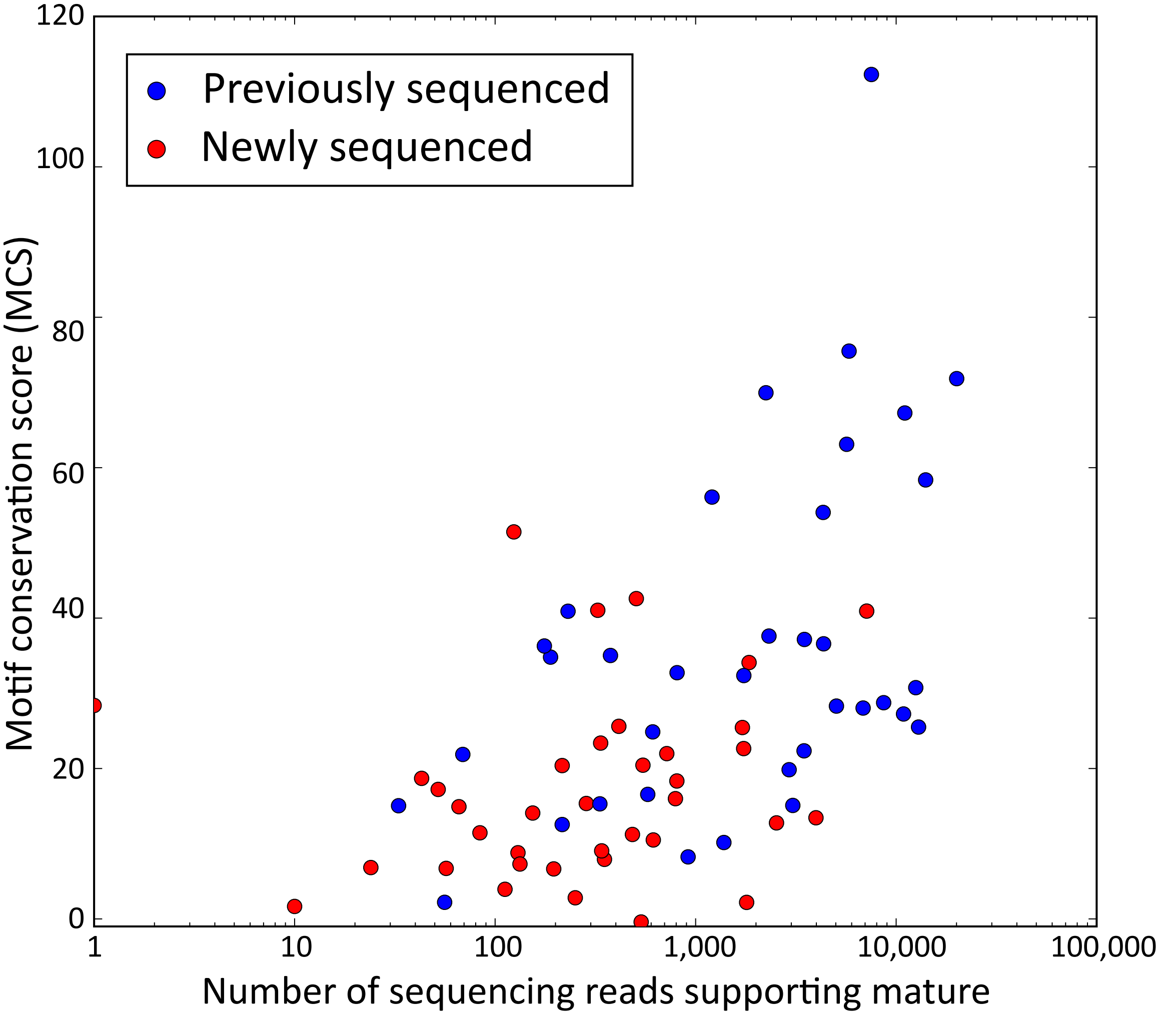


Figure 4-12: Sequencing abundance of miRNAs correlate strongly with the motif conservation score (MCS; Kellis et al., 2003, 2004; Xie et al., 2005) for the corresponding 7-mer in 3′ UTRs of genes. The MCS is the z-score corresponding to the binomial p-value of the fraction of conserved matches of a 7-mer compared to the number expected given other 7-mers of the same composition. As we expect, the more recently sequenced miRNAs (red) have lower expression level. We find a Pearson correlation of 0.53 between MCS and read count. Here and in this section we consider only the family member with the highest expression.

their discovery (high-throughput sequencing). Indeed, this is what we observe: the mean number of mature reads for newly discovered miRNAs is 796 compared to 4337 for the 60 previously identified miRNAs (Mann-Whitney $P = 2.5 \times 10^{-5}$). This follows our performance in hairpin prediction: our top 101 predictions contained 51 of the 60 (85%) previously sequenced miRNAs, but only 33 of the 74 (45%) newly sequenced miRNAs.

Moreover, we find a strong correlation between the MCS score (and consequently the number of $3'$ UTR targets) for each miRNA family and the number of sequenced reads (Figure 4-12). These observations highlight the correspondence between the biological properties (e.g., abundance) and the computational features (e.g., conservation level).

4.5 Conclusion

In this chapter we present two novel algorithms for discovering and characterizing miRNAs. We find that our hairpin prediction algorithm outperforms other comparable algorithms and achieves an unprecedented enrichment for known miRNAs. We also validate most of the predicted miRNAs through intersection with sequencing reads. Because to our knowledge no similar large-scale *de novo* miRNA discovery studies have been released since our original publication of this work, our miRNA discovery algorithm remains state-of-the-art. These methods are broadly applicable and can be used in any circumstance where at least two closely related species and a catalog of known miRNAs is available. The framework is also general and can be extended to include additional diverse features (e.g., chromatin modifications).

4.5.1 Biological contributions

We also make a number of biological contributions. First, we provide a catalog of new miRNAs, including 17 high-confidence predictions with structural and conservation properties consistent with miRNAs. They may represent important, yet condition restricted miRNAs that have eluded sequencing efforts. We also demonstrate the correlation between conservation and biological features: low abundance of a miRNA is correlated with lower levels of conservation and fewer targets. This theme is found elsewhere in this thesis, particularly in the analysis of motif instances.

Chapter 5

Predicting key cell line regulators using chromatin dynamics, regulator expression, and motif enrichments

In this chapter we analyze the relationship between regulatory motif instances and cell line specific annotations of chromatin marks. We will make specific predictions of regulators for human cell lines, seven of which we will test in the next chapter. Some of the analysis presented here was previously published in Ernst et al. (2011) and Consortium et al. (2010), where the experimental data and the creation of chromatin states and their clusters is described. Beyond the generation of this input data, I performed the computational motif analysis presented in this chapter.

5.1 Introduction

The establishment and maintenance of a cell type requires the coordinated action of several transcription factors (Davidson et al., 2003). Moreover, transcription factors have a driving role in cell identity and can be used to reprogram cells to alternate states (Takahashi and Yamanaka, 2006). In this chapter we integrate three datasets in order to predict relevant regulators for cell types: (1) our comparative motif instances; (2) cell line specific chromatin modifications and the corresponding chromatin states; and (3) RNA expression levels for the corresponding cell types.

5.2 Computing enrichments of motifs

A frequent problem in computational biology is the prediction of enriched or depleted motifs within a set of sequences. Very often these sequences will be defined by a list of genes; for example, the promoters of transcripts expressed in a given tissue. In this chapter we consider regions that have similar chromatin modifications, however the same principles apply.

Because of the broad applicability of motif enrichment analysis, it has been extensively studied. Typically, a foreground and background set of sequences is provided with the desire to identify motifs whose matches occur more frequently in the foreground set. In turn, one of two statistics is generally produced for each motif and for both the foreground and background sets: the number of matching sequences (e.g., genes) or the number of motif instances found within each set.

A relatively straightforward approach was taken by Liu et al. (2003): motifs with at least twice as many instances in the promoters of the genes they considered compared to other genes were considered enriched. Typically, however, a simple cutoff on fold-enrichment is avoided and instead a p-value or similar statistical value is computed. For example, Toucan (Aerts et al., 2003) computes an estimate of the number of motif matches for a sequence and compares this to the expected number based on background regions, producing a p-value using a binomial approximation.

ROVER (Haverty et al., 2004) normalizes each motif against background set of regions and then computes a binomial p-value on the number of matching sequences in the foreground. Likewise, Clover (Frith et al., 2004) produces a composite score for each sequence which is then combined across sequences and generates p-values using multiple strategies such as motif or sequence shuffling.

PASTAA (Roider et al., 2009), like Clover, produces a score for each sequence. However, it produces hypergeometric p-values by considering several cutoffs on the sequence scores. McLeay and Bailey (2010) compare several methods for computing enrichment and present linear regression as a competitive way to find enriched motifs.

Some methods also incorporate comparative information to filter motif matches to ones more likely to be functional, but these need to be careful to deal with varying levels of conservation between the foreground and background. Sharan et al. (2003) uses conservation between two species, human and mouse, and computes p-values using a normal approximation correcting for the amount of conservation and length of sequences. oPOSSUM (Ho Sui et al., 2005) also matches motifs to conserved regions, and produces two ranked list based on p-values computed both using a hypergeometric p-value on the number of matching sequences and a binomial p-value correcting for the length and conservation of each sequence.

While finding enriched motifs have been extensively studied, none of the methods we found had all features we desired, particularly: (1) incorporation of motif instances generated using the BLS alignment-free conservation measure (Chapter 2); (2) the production of a meaningful enrichment value that isn't strongly affected by large counts as p-values are; (3) ability to cope with different sizes of input regions, levels of conservation, and compositions.

We address these feature requirements in the following way. Enrichments are computed for each motif and set of regions by computing the fraction of motif instances that fall within the foreground ($r_m = \frac{f_m}{n_m}$, where n_m and f_m indicate the total number of motif instances and the number in the foreground, respectively). This same ratio is computed for control motifs produced as described in Chapter 2. Confidence intervals are then made for each of these ratios (r) using the Wilson score interval (Wilson, 1927) with $z = 1.5$:

$$w(r, n) = \frac{r + \frac{z^2}{2n} \pm z\sqrt{\frac{r(1-r)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Finally, the extreme of each confidence interval is take such that the resulting ratio of original-to-control motif ratios is closest to 1 (if the confidence intervals overlap, the corrected enrichment is 1):

$$e = \begin{cases} w(r_m, n_m)^- / w(r_c, n_c)^+ & \text{if } w(r_m, n_m)^- > w(r_c, n_c)^+ \\ w(r_m, n_m)^+ / w(r_c, n_c)^- & \text{if } w(r_m, n_m)^+ < w(r_c, n_c)^- \\ 1 & \text{otherwise} \end{cases}$$

This corrected enrichment approaches the raw enrichment when counts are very high, but otherwise tends to be conservative and avoids high enrichment or depletion levels due to chance alone.

When precisely two types of regions are considered that partition the total set of regions (e.g., in Figure 5-2), we use a bias-corrected log-odds ratio (Fleiss et al., 2003):

$$l = \log(f_m + 0.5) + \log(n_c - f_c + 0.5) - \log(f_c + 0.5) - \log(n_m - f_m + 0.5)$$

We then use the standard error of this statistic to compute a 95% ($z = 1.96$) confidence interval around each value and again take the most conservative enrichment value:

$$w = z * \sqrt{(f_m + 0.5)^{-1} + (n_c - f_c + 0.5)^{-1} + (f_c + 0.5)^{-1} + (n_m - f_m + 0.5)^{-1}}$$

$$e = \begin{cases} l - w & \text{if } l > w \\ l + w & \text{if } l < -w \\ 0 & \text{otherwise} \end{cases}$$

The benefit of this value is that it is symmetric with respect to which partition of the regions is considered the foreground.

5.3 Predicting cell type specific regulators

We utilize motif enrichment analysis on epigenetic modification-based regions to predict cell line specific regulators and then examine the expression of the corresponding regulators in order to estimate the role of the regulator (either activator or repressor).

Motif enrichment analysis has previously been used on cell type specific data. We performed *de novo* motif discovery and analyzed motif enrichments in five human cell lines, but did not consider the difference between the available cell lines Heintzman et al. (2009). Xi et al. (2007) mapped DNaseI hypersensitivity sites across 1% of the genome in six human cell lines. They found specific known motifs enriched in the hypersensitive sites specific to individual cell lines, but did not correlate these with expression of the corresponding factors. Pennacchio et al. (2007) predicted putative enhancers using conservation and predicted expression cell types using cross validation.

Moreover, several methods have been designed that predict network connections between genes, often using motif instances as inputs (reviewed in Kim et al., 2009). For example, Segal et al. (2008) integrated TF expression along the anterior-posterior axis of the fly to predict expression of genes based on their motif instances. Time courses have also been decomposed in this way, with DREM (Ernst et al., 2007) able to predict TFs responsible for coordinated changes in the expression of genes and then examine the corresponding expression change of driving regulators.

5.4 Individual chromatin marks in human and fly

We start by individually examining human chromatin modification annotations. Specific chromatin modifications are associated with various genomic functional

Modification	Functional association
H3K4me1	Enhancer
H3K4me2	Promoter/enhancer
H3K4me3	Poised/active promoter
H3K27ac	Activation
H3K9ac	Activation
H3K27me3	Repression
H4K20me1	Activation
H3K36me3	Transcription
CTCF	Insulators
WCE	Whole cell extract
RNA	Expression

(a)

Cell line	Source
HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

(b)

Table 5-1: (a) Experiments conducted for each cell line. (b) Encode cell lines used in chromatin analysis.

properties including activation, repression, transcription, enhancers and promoters (Barski et al., 2007; Birney et al., 2007). Moreover, differences in expressed genes across cell types is correlated with corresponding changes in chromatin modifications (Heintzman et al., 2009). Consequently, we reasoned that motifs enriched in cell type specific regions associated with a chromatin mark would be suggestive of function of the corresponding regulators.

As part of the ENCODE project, whole genome annotations of 8 chromatin marks, CTCF, and RNA expression were produced for 9 cell lines (Table 5-1). From this data and for each chromatin modification we define three sets of regions: (1) modified in that cell line (bound; ignoring the state in the other cell types); (2) uniquely modified in the cell line and not bound in any of the remaining cell lines; and (3) uniquely not modified (missing) in the specific cell line, but carrying the modification in all other cell lines.

For example, the motif for NF-κB, an important immune regulator (Baeuerle and Henkel, 1994), is enriched in regions that uniquely have modifications associated with activation in GM12878 (H3K4me1-3, H3K27ac, H3K9ac; Figure 5-1). Conversely, the NF-κB motif is enriched in the regions that do not have do not have the repressive H3K27me3 mark in GM12878, but do in all other cell lines.

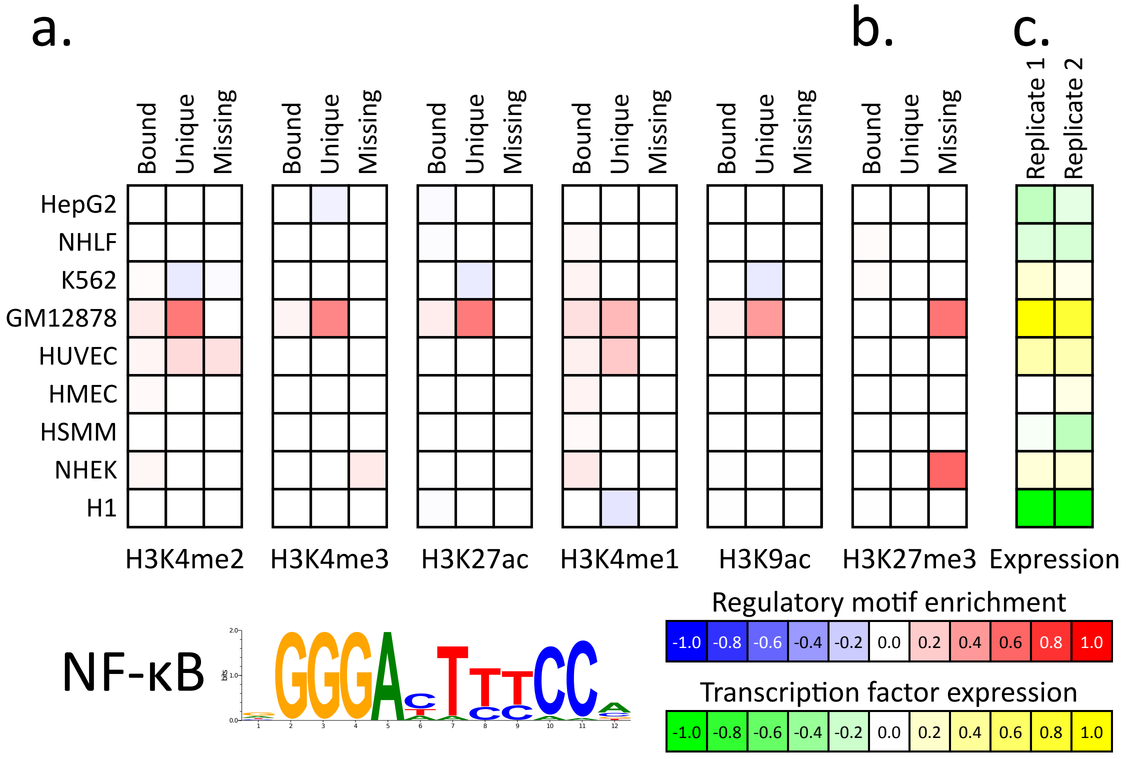


Figure 5-1: Variable motif enrichment is seen for NF-κB in regions of dynamic chromatin. (a) NF-κB motif (matched at at PWM p-value 4^{-6} and 50% confidence) is enriched in active regions unique to GM12878. (b) Conversely, the motif is enriched in regions that are repressed in all other cell types. (c) The NF-κB factor is expressed more highly in GM12878 than other cell lines.

Moreover, we see that NF-κB is expressed particularly highly in GM12878. Together these support a model where NF-κB is expressed in GM12878 and consequently binds to locations leading to their activation. The function of NF-κB in GM12878 has been previously recognized and is required for the establishment of the cell line (Cahir McFarland et al., 1999). Because it seems unlikely that the availability of binding sites for a given factor would result in its expression (although such a model could be developed), this analysis provides a list of potentially functional factors in a cell line.

We also applied similar techniques for predicting regulators that differed between two *D. melanogaster* cell lines, S2 and BG3 (Figure 5-2). Here we examined regions that had the modification uniquely in one cell line versus the other and used the log-odds symmetric enrichment criterion outlined above. Generally we saw that a higher level of enrichment in "active" marks was correlated with lower enrichment in "repressive" marks in the same cell line. Consequently, we defined activators as factors for which expression and enrichment in active marks (or depletion in repressive marks) coincided. For example, CrebA is more highly expressed in S2 cells, and likewise has motif instances that are more enriched in S2 active regions (e.g., H3K4me2) but depleted in S2 inactive regions (e.g., H3K27me3). Twist shows the same trend, but for BG3. We expect to have greater power to predict cell line specific regulators if additional cell lines become available, as they were with human.

5.5 Assessing the activator association of chromatin marks

We reasoned that for factors responsible for the establishment of chromatin marks associated with activation, we expect a positive correlation between the expression vector and the "unique" enrichment. Conversely, for repressive marks we expect a positive correlation between the expression vector and the

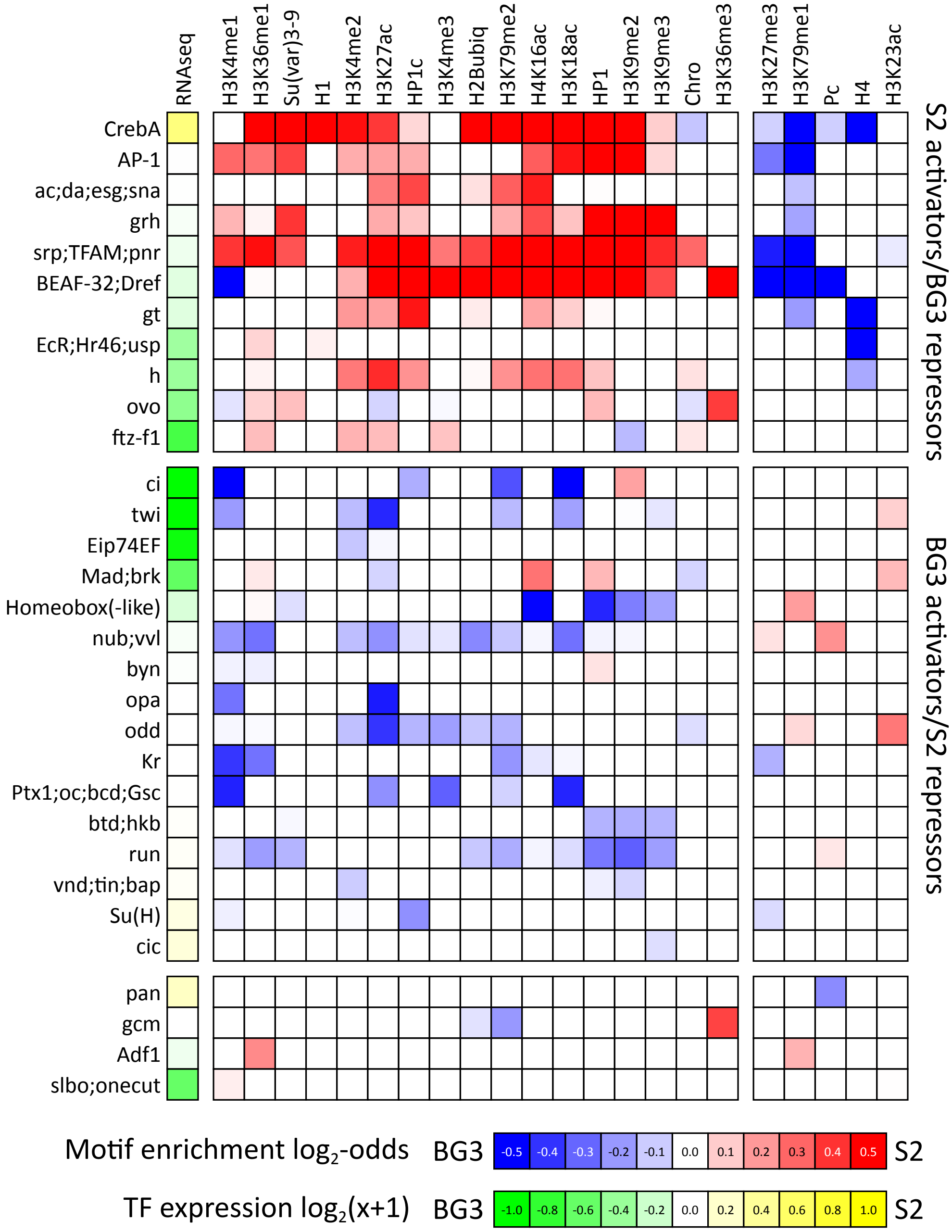


Figure 5-2: Motifs associated with regions of dynamic chromatin modifications. We computed the enrichment comparing S2 (red) and BG3 (blue) of motif instances at a 30% confidence level and 4^{-7} PWM threshold p-value. We consider regions that have the modification in either BG3 or S2, but not both. Enrichments computed using the conservative, bias-corrected log-odds ratio. Motifs for TFs were placed into a cluster if they shared a transcription factor or had a Pearson correlation of at least 0.75 and the motif with the maximal enrichment was used. Ordering of rows and columns above done manually to highlight trends. Modification data from Kharchenko et al. (2011)

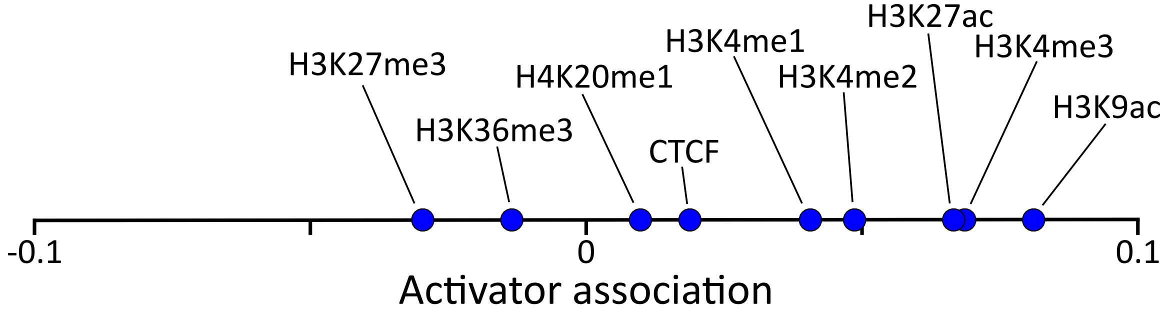
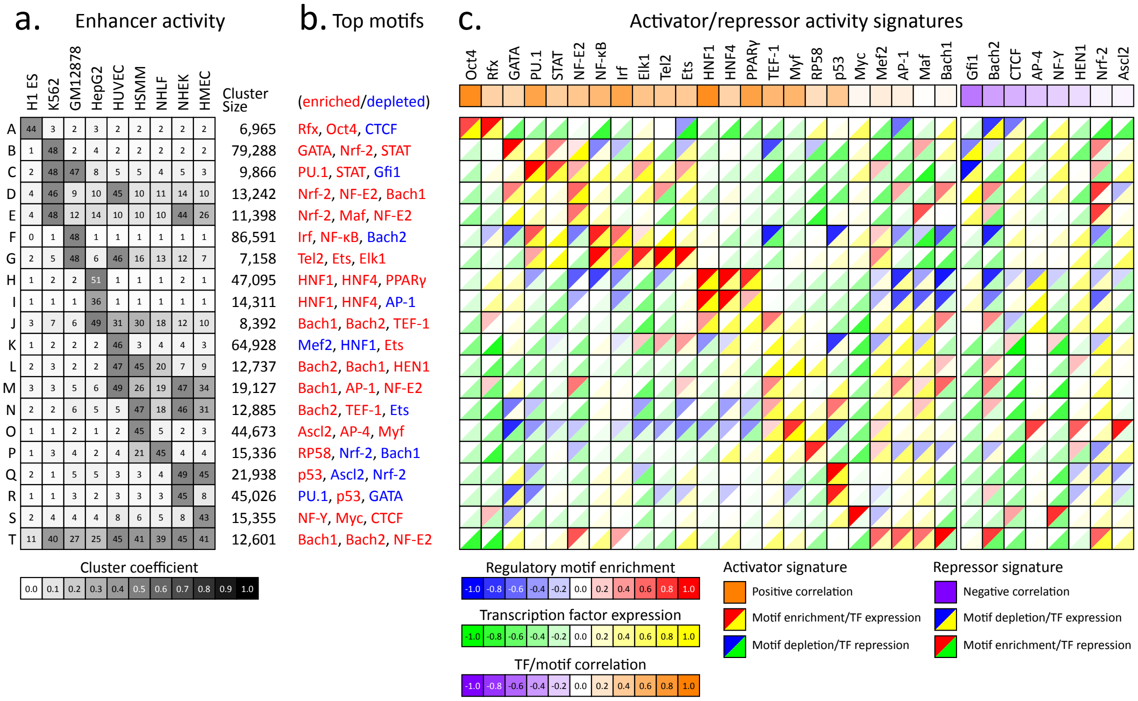


Figure 5-3: Correlation between the enrichment of all activating factor's motif instances in regions unique to a cell type and the expression of the factor in that cell type. We see a positive correlation between enrichment in "activating marks" and expression of factors, while we see a negative correlation for the "repressive mark". Uses expression only in *trans* but recapitulates *cis* association between mark and expression and suggests a casual link between factor binding and chromatin modifications.

"missing" enrichments. To test if this is a global trend, we selected transcription factors in any GO category consistent with activators (GO:0016563, GO:0045944, GO:0045893, GO:0045941, or GO:0003713) that were not also annotated as repressors (GO:0016564, GO:0000122, GO:0045892, GO:0016481, or GO:0003714) and had a known motif. We then computed the correlation between the expression and unique vector, and subtracted the correlation between the expression vector and the missing vector.

Averaging across all factors we found a positive correlation for chromatin modifications associated with activators (H3K4me1-3, H3K27ac, H3K9ac), and negative correlation for H3K27me3 (Figure 5-3). Conversely, for modifications without a clear role in activator or repression (H3K36me3 and CTCF), we found average correlations very close to 0. This remarkable result shows that chromatin modifications can be accurately classified based on the expression of *trans*-activators whose motifs are enriched in them in a cell line specific manner.



a. Enhancer activity

b. Top motifs

	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC	Cluster Size	(enriched/depleted)
A	44	3	2	3	2	2	2	2	2	6,965	Rfx, Oct4, CTCF
B	1	48	2	4	2	2	2	2	2	79,288	GATA, Nrf-2, STAT
C	2	48	47	8	5	5	4	5	3	9,866	PU.1, STAT, Gfi1
D	4	46	9	10	45	10	11	14	10	13,242	Nrf-2, NF-E2, Bach1
E	4	48	12	14	10	10	10	44	26	11,398	Nrf-2, Maf, NF-E2
F	0	1	48	1	1	1	1	1	1	86,591	Irf, NF-κB, Bach2
G	2	5	48	6	46	16	13	12	7	7,158	Tel2, Ets, Elk1
H	1	2	2	51	1	1	1	2	1	47,095	HNF1, HNF4, PPARγ
I	1	1	1	36	1	1	1	1	1	14,311	HNF1, HNF4, AP-1
J	3	7	6	49	31	30	18	12	10	8,392	Bach1, Bach2, TEF-1
K	1	2	2	2	46	3	4	4	3	64,928	Mef2, HNF1, Ets
L	2	3	2	2	47	45	20	7	9	12,737	Bach2, Bach1, HEN1
M	3	3	5	6	49	26	19	47	34	19,127	Bach1, AP-1, NF-E2
N	2	2	6	5	5	47	18	46	31	12,885	Bach2, TEF-1, Ets
O	1	1	2	1	2	45	5	2	3	44,673	Ascl2, AP-4, Myf
P	1	3	4	2	4	21	45	4	4	15,336	RP58, Nrf-2, Bach1
Q	2	1	5	5	3	3	4	49	45	21,938	p53, Ascl2, Nrf-2
R	1	1	3	2	3	3	3	45	8	45,026	PU.1, p53, GATA
S	2	4	4	4	8	6	5	8	43	15,355	NF-Y, Myc, CTCF
T	11	40	27	25	45	41	39	45	41	12,601	Bach1, Bach2, NF-E2

Cluster coefficient

0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Figure 5-4: Characterization of clustered enhancer activity patterns (a) For each cluster, the average weight of the strong enhancer State 4 (percent) in each of the cell types. The number of 200bp windows that fall into each cluster. (b) Top most enriched (red) and depleted (blue) motifs in each of the clusters. (c) Top left of each box: enrichment ($\log_2$) of all motifs from panel (b) shown across all clusters. Bottom right of each box: expression in each cluster for the corresponding factor. Correlation between these two vectors shown on top in the purple/orange scale.

5.6 Chromatin state analysis predicts cell line regulators

While looking at marks in isolation allowed us to predict cell line specific regulators, it is recognized that individual marks often do not capture the unique state of a region and moreover several marks can be redundant (Kharchenko et al., 2011; Ernst and Kellis, 2010). Further, the uniqueness criteria can be restrictive because many regulators may act in two or more cell types. Consequently, we replaced our use of a single chromatin mark to define regions with a combination of marks consistent with an enhancer chromatin state, as provided in Ernst et al. (2011). Further, rather than look only at regions that were uniquely on (or off) in a single cell type, we partitioned the enhancer regions into 20 clusters of different on/off patterns (using k-means). In addition to motif enrichments, and as we did for individual marks, we incorporated expression values for the factors (averaging over multiple factors recognizing the same motif, when applicable). However, because a cluster of cell types does not have an expression, we instead used the Pearson correlation between the expression vector and the cluster center.

We compared motif enrichments across these clusters (using as the background the union of all enhancers) suggesting several cell-type specific regulators (Figure 5-4). For example, we found the Oct4 motif enriched in enhancers specific to human embryonic stem cells and found a coordinated expression of Oct4, consistent with the role Oct4 plays in embryonic stem cells (Loh et al., 2006). Moreover, for HNF1 and HNF4 we found enrichment and expression in HepG2 (liver carcinoma), consistent their importance in liver function and development (Costa et al., 2003).

We also see enrichments that are present in clusters across several cell lines. Ets factors are a predicted regulator for enhancers that are active in both GM12878 and HUVEC (cluster G), but not those only active in either one alone (Clusters F and K). p53 has a widespread role in regulation (Wei et al., 2006) and is found as an activator for four primary cell lines, HSMM, NHLF, NHEK and HMEC (clusters N, Q, and R), but not in cell lines where it is inactivated due to mutation (K562; Law et al., 1993), viral infection (GM12878; Forte and Luftig, 2009), or cytoplasmic localization (ES; Solozobova et al., 2009).

For these factors (and several others) we saw a correlation between expression and enrichment across the 20 clusters, which we model as a signature of an activator. We see the opposite trend for putative repressors: Gfi1 and Bach2 show an anti-correlation between expression and enrichment and indeed these two factors have known repressive roles (Muto et al., 1998; Hock and Orkin, 2006). For some factors we see minimal correlation, which may be due to us not having the correct specific gene tied to the motif or due to post transcriptional regulation of TF activity.

5.7 Conclusion

In this chapter we show how motif enrichments coupled with expression analysis can predict putative regulators. We also show evidence that motifs are the drivers of chromatin modifications and that the modifications can be accurately classified by their trans relationships with factors. Because it seems unlikely that the accessibility of binding sites leads to the expression of factors, the results of this chapter support a model where expression of a factor leads to the establishment of appropriate chromatin domains where it binds.

The predictions we make in this chapter are a compelling biological contribution on their own, particularly in the context of their agreement with known roles of the specific factors. However, we will systematically test seven of our predictions in the next chapter.

Chapter 6

Systematic design and testing of enhancer sequences for dissection of regulatory motif function

In this chapter we test the predictions of Chapters 2 and 5 by selecting predicted regulators for two cell lines and systematically, interrogating their expression. The chromatin states used in this chapter were provided prior to publication by Jason Ernst and the experimental work was carried out by Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, and Tarjei Mikkelsen. The experimental technique we use in this chapter prior to its publication was developed by Alexandre Melnikov, Peter Rogov, Anand Murugan, Xiaolan Zhang, and Tarjei Mikkelsen. I chose and manipulated the putative enhancer sequences and did the subsequent analysis from the raw read counts.

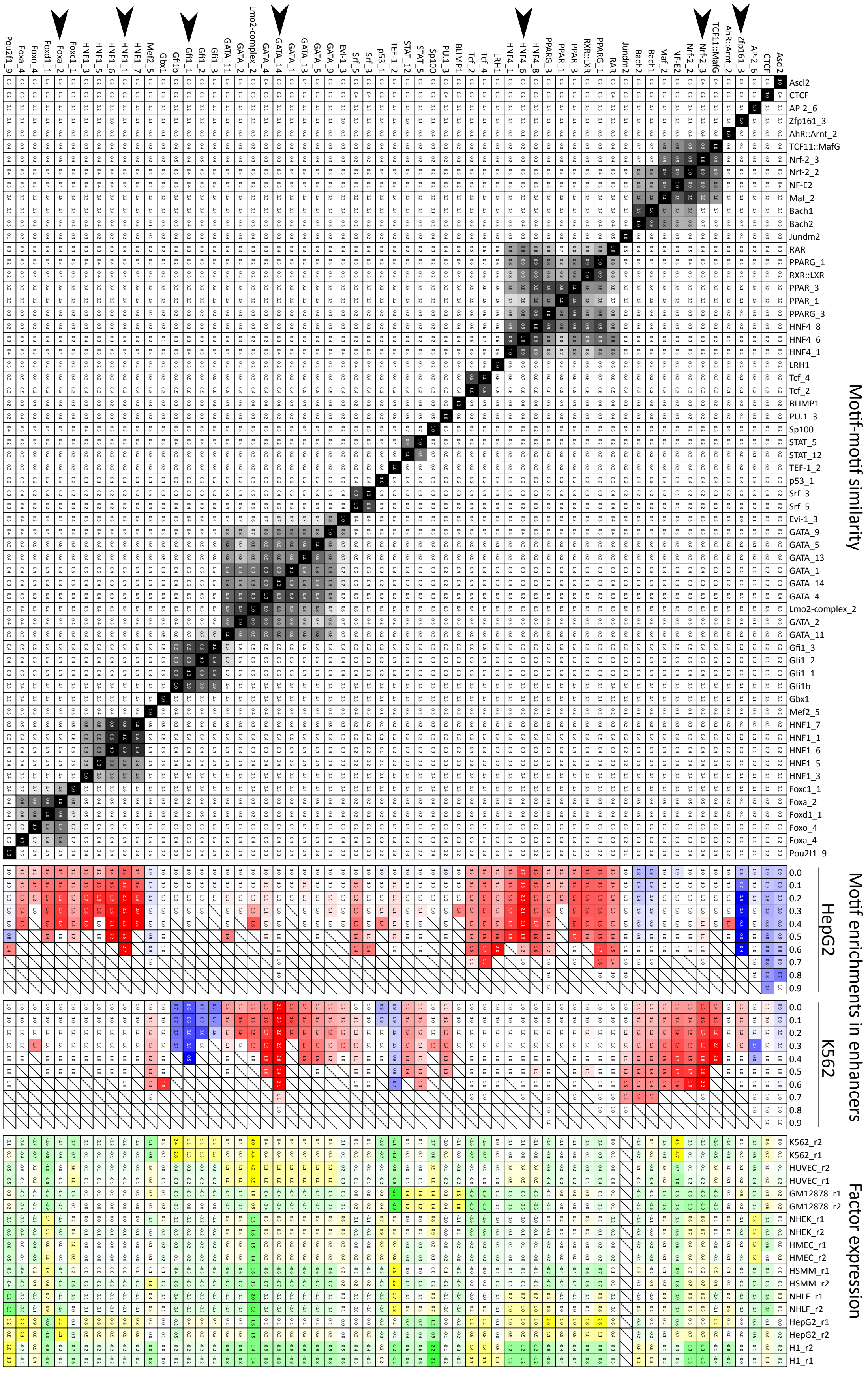
6.1 Introduction

The recent proliferation of ChIP-seq applied to chromatin modifications has permitted the large-scale prediction of putative enhancer regions in the genome (Ernst and Kellis, 2010; Heintzman et al., 2009). However, previous approaches to large scale identification of the specific factors responsible for establishing the enhancers (e.g., ChIP-chip or Chip-seq) have been either low resolution and unable to show the necessity of binding (Ren et al., 2000; Iyer et al., 2001; Robertson et al., 2007) or focused on only a few sites (Patwardhan et al., 2009; Baliga, 2001). Further, while evolutionarily conserved motif instances for transcription factors (TFs) make highly precise predictions for the nucleotides involved in regulation (Chapter 2; Moses et al., 2004; Xie et al., 2005; Kheradpour et al., 2007), their large scale validation has not been previously attempted.

In this chapter we present the systematic testing of two major aspects of this thesis: the motif instance predictions made in Chapter 2 and the relationships between specific motifs and cell lines made in Chapter 5. We employ a massively parallel reporter assay (MPRA; Melnikov et al., 2012) to measure the transcriptional levels induced by hundreds of erythroleukemia (K562) and liver carcinoma (HepG2) enhancers (Ernst et al., 2011) centered on the regulatory motif instances.

For each of 2,104 instances of five activator and two repressor motifs, we made directed modifications of the bases corresponding to these motif matches (while keeping the flanking bases constant) and observe the change in expression with 10 unique bar-codes for each of 5,418 variants, resulting in 54,180 distinct expression measurements. For TFs with the signature of activators, we find robust evidence that: (1) enhancers centered on comparative motif instances have higher expres-

Figure 6-1 *(on the next page)*: Analysis of all motifs that are enriched in enhancers for HepG2 or K562 (relevant details in Figure 6-2). (a) The motif-motif correlation (at all shifts including reverse complement). Different variants of literature motifs are numbered. (b) The enrichment (red) or depletion (blue) of the motif at different comparative confidence cutoffs in enhancers for the indicated cell line. (c) The relative expression ($\log_2$) of the factor corresponding to the motif in the cell lines. Where several proteins exist for a given motif we average the expression values. We ultimately choose (indicated in order) Zfp161_3 (Badis et al., 2009), Nrf-2_3 (Jaspar MA0150.1), HNF4_6 (Transfac M00158), GATA_14 (Jaspar MA0140.1), Gfi1_1 (Transfac M00250), HNF1_1 (Jaspar MA0046.1), and Foxa_2 (Jaspar MA0047.2) on the basis of their enrichment/depletion, expression and sequence-level uniqueness. We note that because our analysis is unable to distinguish between factors which share a motif, we may not capture the expression of all potential alternative factors.


Motif-motif similarity
Motif enrichments in enhancers
HepG2
K562
Factor expression
Ascl2
CTCF
AP-2_6
Zfp161_3
AhR::Arnt_2
TCF11::MafG
Nrf-2_3
Nrf-2_2
NF-E2
Maf_2
Bach1
Bach2
Jundm2
RAR
PPARG_1
RXR::LXR
PPAR_3
PPAR_1
PPARG_3
HNF4_8
HNF4_6
HNF4_1
LRH1
Tcf_4
Tcf_2
BLIMP1
PU.1_3
Sp100
STAT_5
STAT_12
TEF-1_2
p53_1
Srf_3
Srf_5
Evi-1_3
GATA_9
GATA_5
GATA_13
GATA_1
GATA_14
GATA_4
Lmo2-complex_2
GATA_2
GATA_11
Gfi1_3
Gfi1_2
Gfi1_1
Gfi1b
Gbx1
Mef2_5
HNF1_7
HNF1_1
HNF1_6
HNF1_5
HNF1_3
Foxc1_1
Foxa_2
Foxd1_1
Foxo_4
Foxa_4
Pou2f1_9
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
K562_r2
K562_r1
HUVEC_r2
HUVEC_r1
GM12878_r1
GM12878_r2
NHEK_r1
NHEK_r2
HMEC_r1
HMEC_r2
HSMM_r1
HSMM_r2
NHLF_r1
NHLF_r2
HepG2_r1
HepG2_r2
H1_r2
H1_r1

sion than those centered on motif instances ignoring conservation, (2) scrambling or removing the motif instance results in a significant reduction in expression and (3) mutations that do not disrupt the motif consensus have no effect on expression. Conversely, disrupting motif instances corresponding to a repressive TF can lead to an increase in expression in a cell type where the enhancers are usually not active.

Lastly, we identify additional characteristics of these enhancers that show highest wild-type activity and find that signatures of nucleosome exclusion and comparative motif information are most informative. These results strongly confirm the cell line specific enhancer predictions, the role of evolutionary signatures for motif instance prediction, and that MPRA is a viable technique for measuring enhancer activity, suggesting a general strategy for deciphering *cis*-regulatory elements. Further, this work provides as a resource thousands of tested enhancers, hundreds of which we estimate to be functional. This is comparable to the number of enhancers tested in mammalian systems *in vivo* (Visel et al., 2007) and is, to our knowledge, the largest dataset in human cell lines.

6.2 Enhancer sequence selection and design

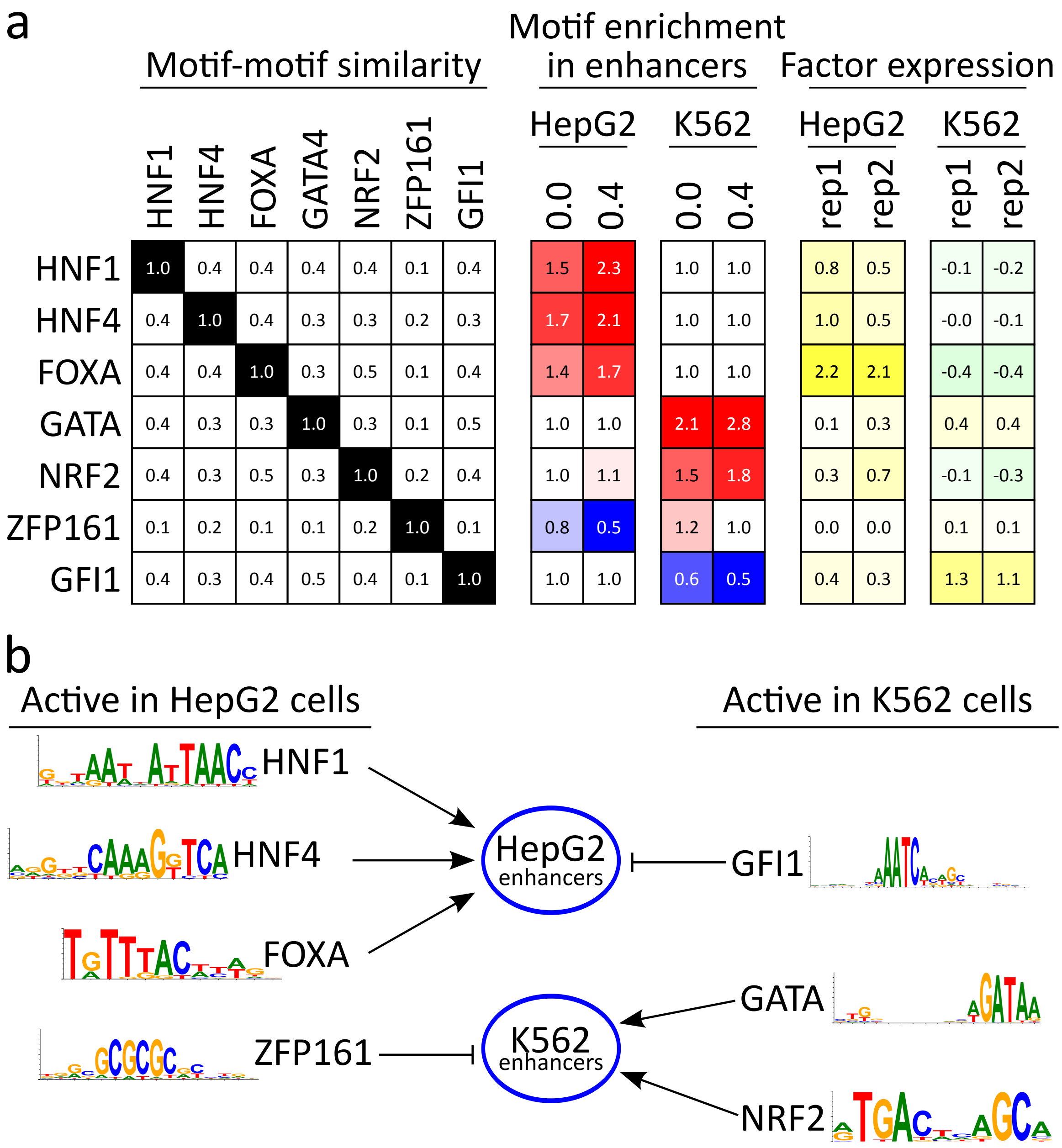
We define cell line specific enhancers as the union of states 4 and 5 ("strong enhancers") from (Ernst et al., 2011) excluding regions within 2kb of a TSS and select two cell lines, HepG2 (liver carcinoma) and K562 (erythrocytic leukemia), for in experimental validation. We choose these enhancers because they are produced for cell lines for which MPRA is amendable and were available at the start of this study. We use our human motif instances described in Chapter 2 and employ the activator and repressor signatures defined in Chapter 5 to systematically select relevant motifs for these two cell lines (Figure 6-1). This results in 3 putative activators for HepG2: HNF1, HNF4 and FOXA, all involved in liver development and function (Costa et al., 2003), and two activators for K562: the hematopoiesis regulator family GATA (Weiss and Orkin, 1995) and NRF2 (Figure 6-2). We also select ZFP161 and GFI1 as repressors for HepG2 and K562, respectively, based on depletion of motifs in the corresponding cell line and note that both these factors have previously been recognized as repressors (Sobek-Klocke et al., 1997; Hock and Orkin, 2006).

The technology we employ (MPRA) permits the testing the enhancer activity of 145-bp sequences. We select wild-type sequences from the genome by taking matches to the selected motifs that fall within cell-line specific enhancers, with the specific number selected described in Table 6-1. We test each wild-type putative enhancer along with several directed manipulations in order to verify that transcription factor binding was responsible for expression.

For every tested wild-type sequence we generate a version with only the motif match scrambled. The specific permutation used for each motif is determined by creating 100 random scrambles and choosing the one with the lowest similarity to the original motif (Figure 6-3). For each TF we also select 15 sequences (Table 6-1) and produce an additional 6 manipulations: (1) complete removal of the motif match, with additional flanking sequence to fill the 145-bp; the single base pair change that (2) maximally reduces, (3) makes the smallest change or (4) maximally increases the PWM match score; and (5,6) two random manipulations performed by choosing two positions inside the motif match (without replacement) and changing them to one of the other 3 bases regardless of the effect it has on the PWM match score.

6.3 Experimental enhancer activity determination

Experimental measurement of enhancer activity for all generated sequences is performed in both K562 and HepG2, regardless of the cell line where the enhancer was originally found. We employ a massively parallel reporter assay (MPRA) for this purpose (details in Melnikov et al., 2012) with two biological replicates and 10 unique bar codes for each tested sequence (overview in Figure 6-4). Briefly, oligonucleotide libraries were synthesized by Agilent, Inc. containing, in order,



	Motif-motif similarity							Motif enrichment in enhancers				Factor expression			
								HepG2		K562		HepG2		K562	
	HNF1	HNF4	FOXA	GATA4	NRF2	ZFP161	GFI1	0.0	0.4	0.0	0.4	rep1	rep2	rep1	rep2
HNF1	1.0	0.4	0.4	0.4	0.4	0.1	0.4	1.5	2.3	1.0	1.0	0.8	0.5	-0.1	-0.2
HNF4	0.4	1.0	0.4	0.3	0.3	0.2	0.3	1.7	2.1	1.0	1.0	1.0	0.5	-0.0	-0.1
FOXA	0.4	0.4	1.0	0.3	0.5	0.1	0.4	1.4	1.7	1.0	1.0	2.2	2.1	-0.4	-0.4
GATA	0.4	0.3	0.3	1.0	0.3	0.1	0.5	1.0	1.0	2.1	2.8	0.1	0.3	0.4	0.4
NRF2	0.4	0.3	0.5	0.3	1.0	0.2	0.4	1.0	1.1	1.5	1.8	0.3	0.7	-0.1	-0.3
ZFP161	0.1	0.2	0.1	0.1	0.2	1.0	0.1	0.8	0.5	1.2	1.0	0.0	0.0	0.1	0.1
GFI1	0.4	0.3	0.4	0.5	0.4	0.1	1.0	1.0	1.0	0.6	0.5	0.4	0.3	1.3	1.1

Figure 6-2: The motifs ultimately chosen for analysis. (a) Selected rows and columns from Figure 6-1. Activators and repressors for HepG2 and K562 were chosen based on the three criteria: (1) minimal motif-motif similarity, (2) enrichment for activators and depletion for repressors in the cell line of interest and (3) expression of the corresponding factor in the target cell line. (b) Diagram showing selected activators and repressors. We show that by manipulating the motif matches we are able to produce a cell line specific reduction in expression for the activators.

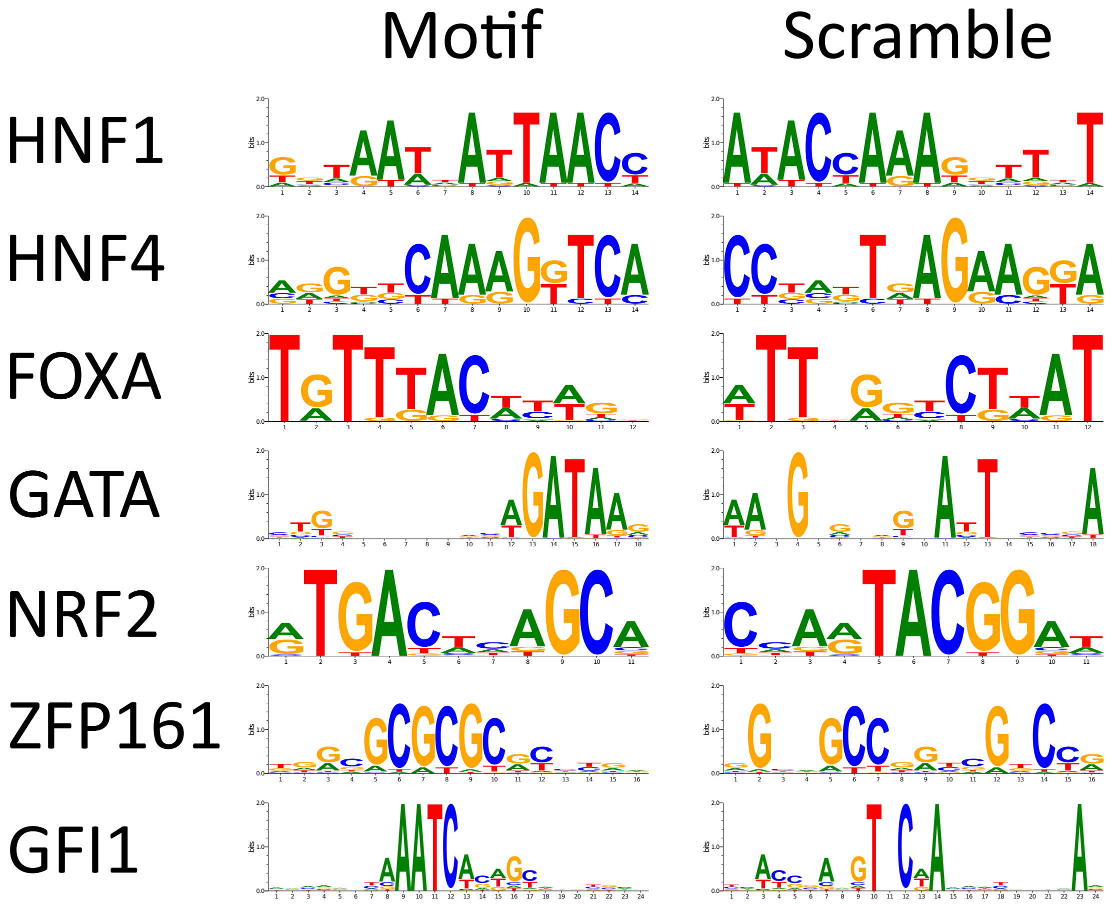


Figure 6-3: Logos of the specific motifs selected for analysis and the permutation order used for scrambling applied to the matrix.

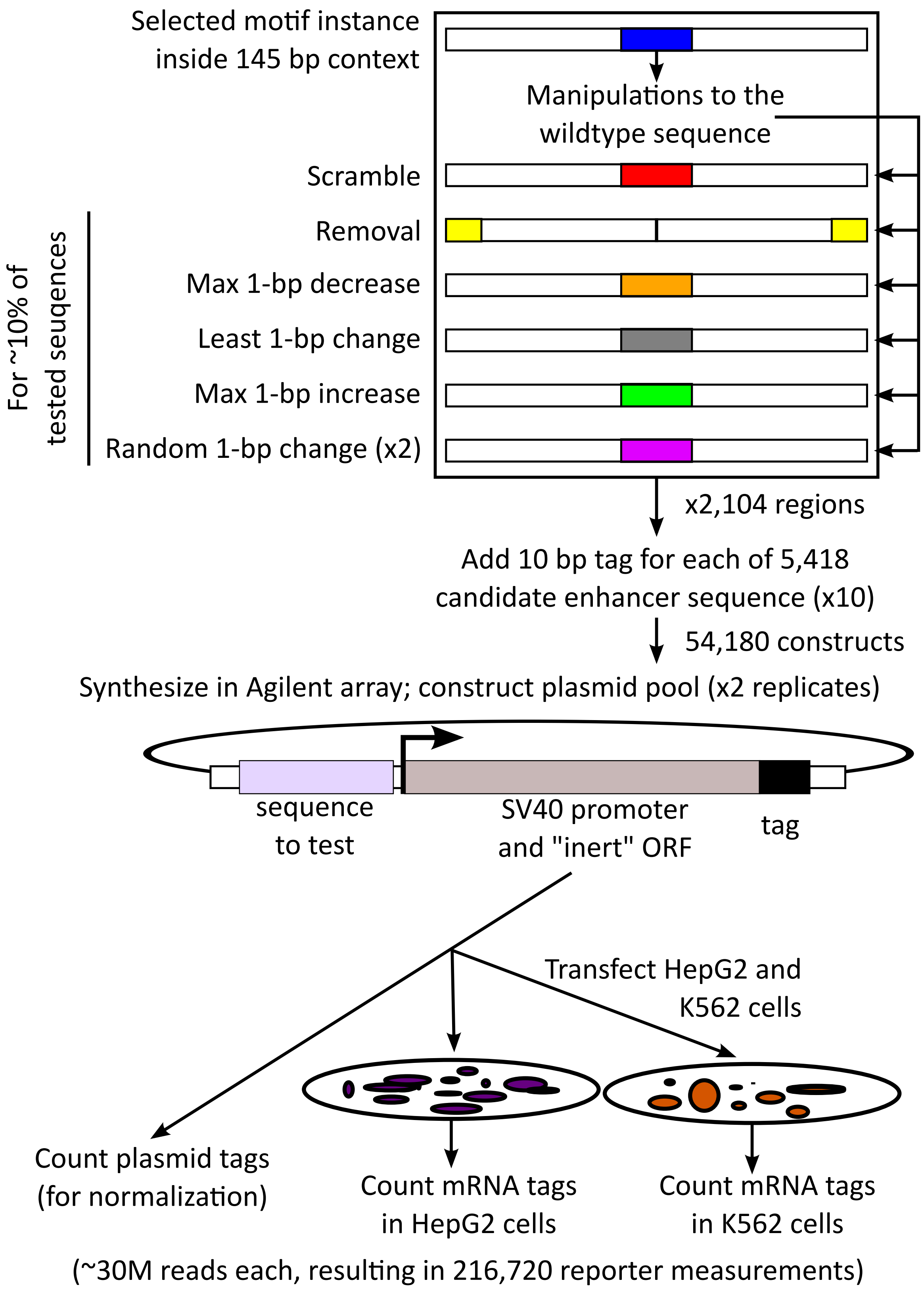


Figure 6-4: Overview of computational and experimental procedure used to test motif instances found within enhancers.

	Activators	Repressors
HepG2	HNF1, HNF4, FOXA	ZFP161
K562	GATA, NRF2	GFI1
Same cell line	160	18
+scramble	160	18
+other manipulations	15 (x6)	0
Opposite cell line	18	160
+scramble	18	160
+other manipulations	0	15 (x6)

Table 6-1: Number of tested sequences for each class and factor. This design was repeated twice; once for comparative instances (Chapter 2) and once for motif matches ignoring conservation (which could overlap the comparative instances). Some sequences were not included for technical reasons or due to too few motif matches; see Table 6-2. For the comparative instances, we always choose the instances with highest confidence. Ties are broken randomly.

the tested sequence, KpnI/XbaI restriction sites, and a variable 10-bp tag sequence and flanking prime sites (a total of 200-bp). Full-length oligonucleotides were isolated (Visel et al., 2009) and directionally cloned into the MPRA vector (Patwardhan et al., 2009). The KpnI/XbaI restriction sites were used to insert SV40 promoter and the luc2 ORF5 into the constructs. Transfections were performed into 5×10^6 HepG2 cells (using FugeneTM HD) and into 4×10^6 K562 cells (using Nucleofection).

The plasmid, K562 mRNA, and HepG2 mRNA libraries were sequenced using 36-nt single-end reads on the Illumina HiSeq 2000 instrument. Reads matching one of the 54,000 designed tags were counted and divided by the total for each pool. Each mRNA count was then divided by the corresponding plasmid count (excluding tags with fewer than 40 plasmid reads) and the $\log_2$ ratio, divided by the median for that cell line, was used as our expression value (where 0 is taken as the baseline in our plots). However, because only a small portion of our tested sequences corresponded to what we later determined to be a functional enhancer for each cell line, we estimate that the 0 baseline is approximately the background level of expression for our promoter. We find that the data is reproducible (Figure 6-5).

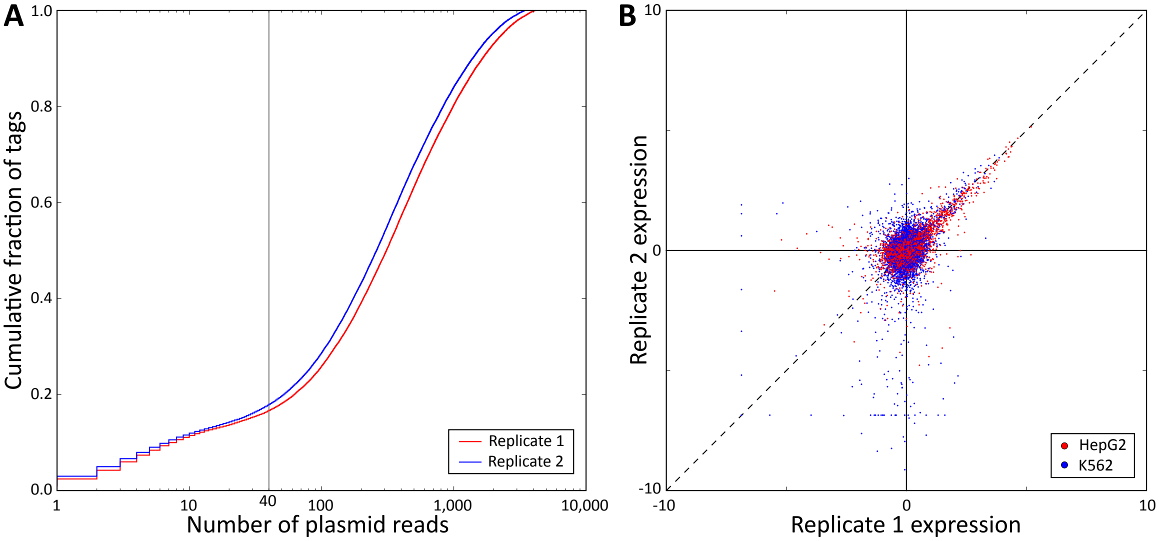


Figure 6-5: (a) Cumulative distribution of the plasmid read counts for the two replicates. 17.6% and 16.5% of tags for replicates 1 and 2, respectively did not pass the criteria of having at least 40 reads. However, because each tested sequence was tested with 10 unique tags per replicate, only 46 of 5,418 (0.85%) tested sequences had no passing tags in both replicates. (a) Comparison of the two replicates using the same averaging procedure as was employed when coming all data (see Methods). Pearson correlation is 0.69 and 0.36 for HepG2 and K562 data, respectively. The lower correlation seen for K562 is attributable to biological properties of the cell line and the transfection procedure. These correlations are within the 95% confidence intervals when comparing tags 1-5 to 6-10 across both replicates (0.63-0.69, and 0.27-0.36, respectively), indicating we cannot attribute a significant portion of the difference to biological variance. 60 of the 5,418 tested sequences were excluded due to not having any plasmids with sufficient (40) plasmid reads in at least one replicate.

6.4 Results

6.4.1 Comparative motif instances for activators select functional enhancers

Figure 6-6 shows the results for a tested HepG2 enhancer centered on a alignment-free conserved (comparative) HNF4 motif instance. We see that the tested region falls within a 'dip' in the H3K27ac chromatin signal (Figure 6-6b). Expression is completely absent in K562 (Figure 6-6c) where we predict neither HNF4 nor the enhancer to be active. However we see robust expression for the original sequence in HepG2, which is abolished by any of the disruptive mutations but maintained by the non-disruptive mutations.

These trends generalize to the other tested comparative HNF4 instances in HepG2 enhancers (Figure 6-7) and beyond to the other activator / cell-line combinations (Figure 6-8 and 6-17). For all activators the expression level driven by sequences selected using comparative motif matches significantly drops when the motif match is scrambled (combined Wilcoxon $P_W = 3 \times 10^{-54}$). Moreover, the expression level of the comparative enhancers with scrambled motifs is lower than that of random motif matches (combined Mann-Whitney $P_U = 7.6 \times 10^{-10}$). This drop in expression to background levels demonstrates that the motif matches are necessary for the enhancer activity of these sequences in the assayed cell line.

We tested additional disruptive mutations for 74 comparative activator motif instances. For these 74 loci, all three disruptive mutations (scramble, removal, max 1-bp decrease) resulted in significantly reduced expression levels (combined $P_W = 2.2 \times 10^{-7}$, 1.5×10^{-4} and 1.7×10^{-6}, respectively). Moreover, we did not find a statistically significant difference in the expression between scrambling and complete removal of the motif match ($P_W = 0.1335$) and scrambling resulted in only a slightly significantly lower expression level than than the best 1-bp reduction ($P_W = 0.0454$). We found no significant increase or decrease in the resulting expression for the 1-bp neutral modification ($P_W = 0.0814$), however, there was

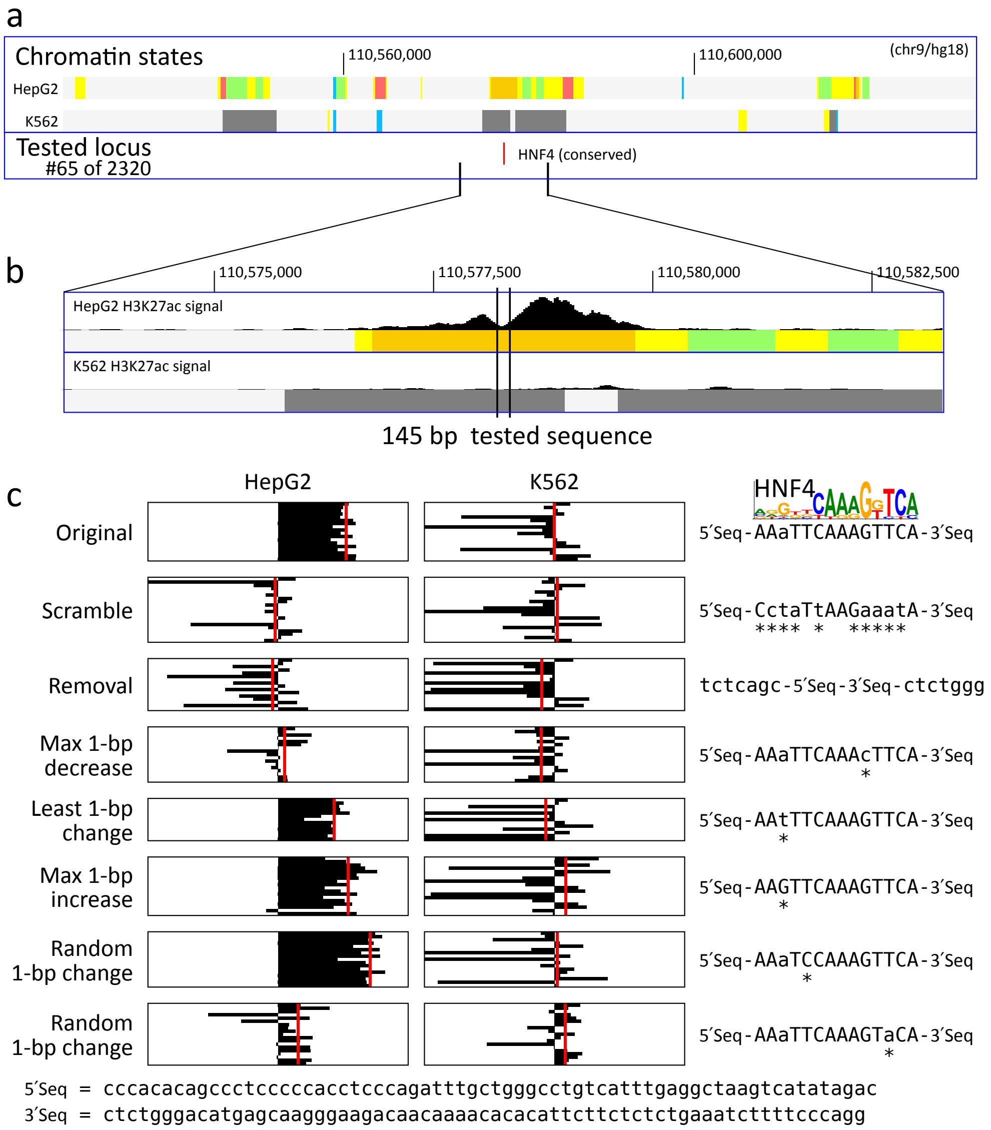


Figure 6-6: Example putative enhancer region centered on a motif match for an activator inside a HepG2 enhancer. (a) The tested sequence was selected for being centered on a HNF4 motif match and found in a HepG2 enhancer. Chromatin state tracks are colored with promoters in red, poised promoters in purple, enhancers in orange (strong) and yellow (weak), insulators in blue, transcribed states in green, repressed regions in grey and low signal/repetitive regions in light grey (described in Ernst et al., 2011). (b) The H3K27ac signal shows a significant dip in HepG2 coincident with the tested region, a signature of potential nucleosome exclusion consistent with the binding of a TF. (c) The original sequence shows strong expression in HepG2 (replicates shown as black bars, combined value shown as red line) whereas expression is absent in K562. Disruptive mutations (scramble, removal and max 1-bp decrease) eliminate expression, whereas the neutral and particularly the max 1-bp increase do not. Random mutations have behavior consistent with whether mutation is tolerated by the motif consensus (upper case) or not (lower case). Bases changed relative to original sequence are indicated by stars (*).

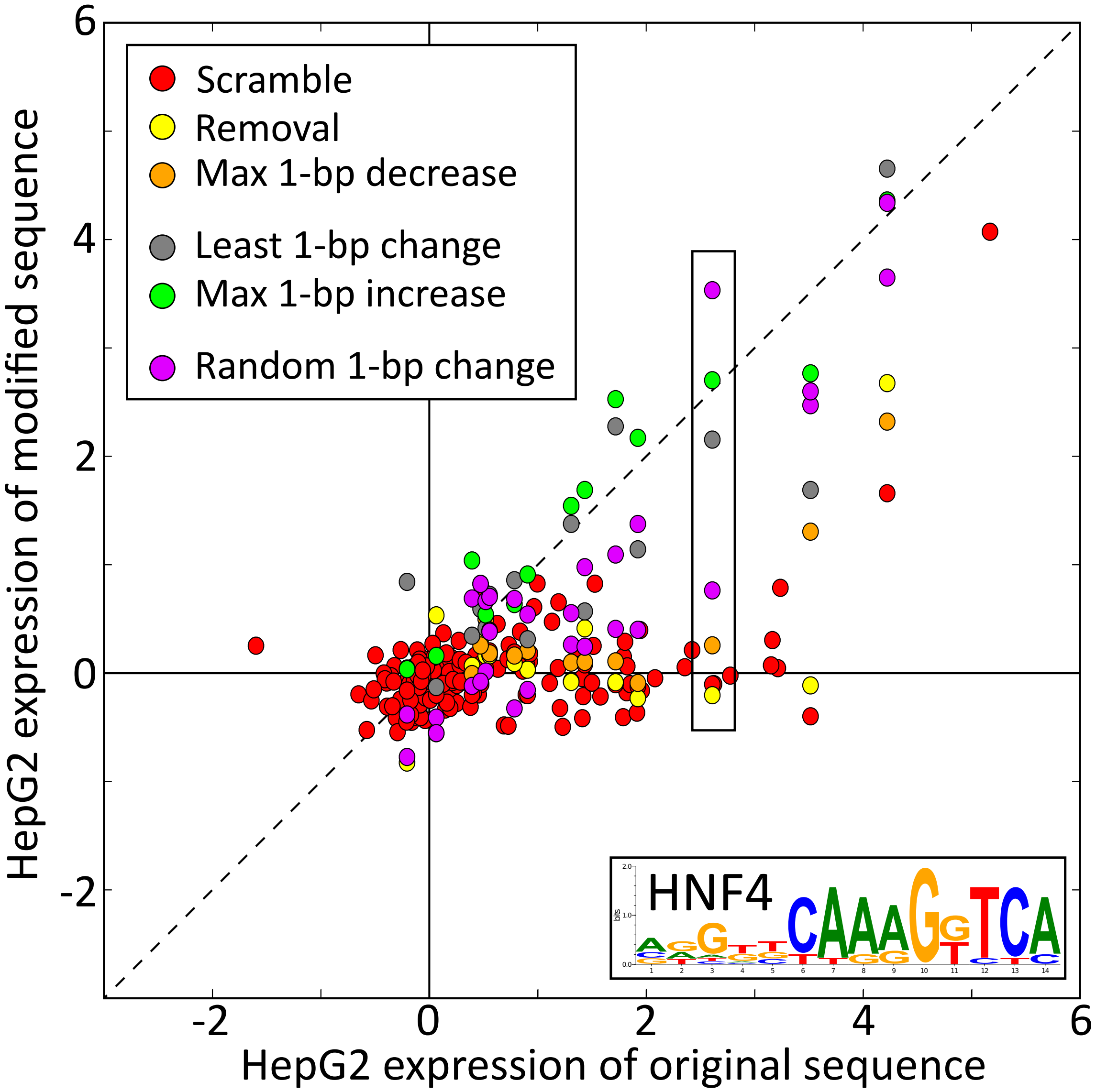


Figure 6-7: Average reporter gene expression for each of the 160 predicted HepG2 enhancers centered on comparative HNF4 motif instances. Manipulations featured in Figure 6-6 are boxed. The original (wild-type) expression is shown on the x-axis, and the engineered construct is shown on the y-axis. The 160 scramble manipulations stay largely on the x-axis, demonstrating the ability of the manipulation to disrupt expression. A similar trend is seen for the other two disruptive manipulations (greatest 1-bp decrease in orange and removal in yellow). The two non-disruptive mutations (neutral in gray; match increasing in blue) are largely are unaffected by the mutation and random (purple) shows a intermediate change.

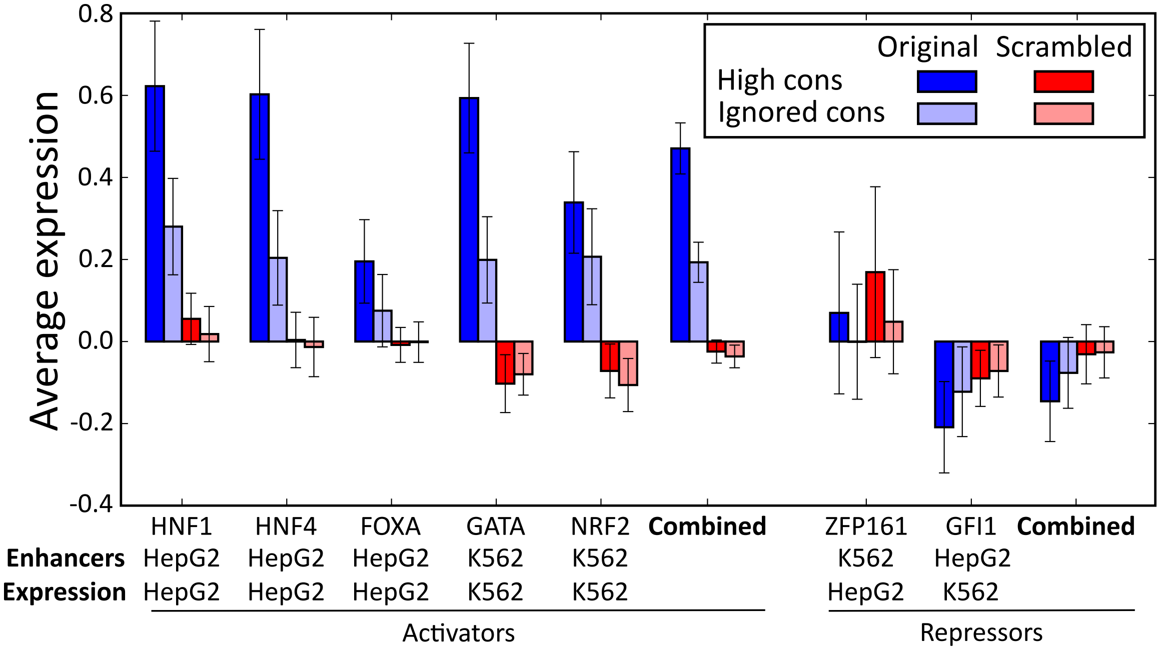


Figure 6-8: Comparison of the effect of the scramble manipulation for enhancers centered on both alignment-free conserved and all motif instances. Corresponding p-values are in Figure 6-9.

a significant increase ($P_W = 0.0056$) in expression for the 1-bp match score improvement. The consistent difference between the 1-bp changes that reduce the motif match strength versus those that do not demonstrate that the specificity to the motif sequence is the driving aspect responsible for the observed changes in enhancer activity. For the random manipulation, we found a significant correlation between the reduction in expression and the match score difference for those

Figure 6-9 *(on the next page)*: (a) P-values showing a consistent and strongly significant reduction in expression specific to disruption of motif binding sites. P-values are computed using the ^{a}Wilcoxon signed-rank and ^{b}Mann-Whitney U two-tailed non-parametric tests. Headers indicate the maximum number of tested sequences per factor in parenthesis (precise numbers are in Table 2). P-values in parenthesis indicate an increased value in either the ^{a}instances ignoring conservation or for the ^{b}modified sequence. Values used in in the text are highlighted in bold. Corresponding to bar plots in Figures 6-8 and 6-17a. (b) Same as (a), except with expression cell line reversed to demonstrate cell type specificity. A notable exception is NRF2. (c) Same as (a), except with both expression and enhancer cell lines switched. Only 18 such sites were selected per factor, and they were only tested with the scramble modification. Corresponding to bar plots in Figure 6-17b.

a

		Enhancer cell line	Expression cell line	High vs. ignored conserv. (160)[a]	High conservation: Scramble (160)[b]	Removal (15)[b]	Max 1-bp decrease (15)[b]	Least 1-bp change (15)[b]	Max 1-bp increase (15)[b]	Random 1-bp change (30)[b]	Ignored conservation: Scramble (160)[b]	Removal (15)[b]	Max 1-bp decrease (15)[b]	Least 1-bp change (15)[b]	Max 1-bp increase (15)[b]	Random 1-bp change (30)[b]
Activators	HNF1	HepG2	HepG2	2.3×10^{-4}	1.7×10^{-13}	0.0409	0.0038	0.0146	(0.0783)	8.9×10^{-5}	1.9×10^{-6}	0.8647	(0.7333)	(0.6496)	(0.0199)	(0.5038)
	HNF4	HepG2	HepG2	2.1×10^{-4}	8.0×10^{-15}	0.0018	0.0012	0.4955	(0.0609)	1.3×10^{-4}	4.4×10^{-4}	(0.9096)	0.9096	0.9096	(0.0090)	(0.8130)
	FOXA	HepG2	HepG2	0.1618	**0.0035**	(0.8647)	(0.7764)	(0.4955)	(0.1398)	(0.7343)	0.5947	0.0736	0.3003	0.7299	0.4326	0.3164
	GATA	K562	K562	1.7×10^{-6}	1.9×10^{-18}	0.1118	0.1398	(0.9547)	0.9547	0.0196	9.3×10^{-6}	0.4265	0.7333	(0.4955)	(0.0054)	(0.5857)
	NRF2	K562	K562	0.0352	4.3×10^{-11}	0.4326	0.0303	0.4326	(0.2209)	0.3745	6.3×10^{-7}	0.3739	0.1307	0.5937	(0.1823)	0.0015
	Combined			**9.0×10^{-13}**	**2.8×10^{-54}**	**1.5×10^{-4}**	**1.7×10^{-6}**	**0.0814**	**(0.0056)**	1.6×10^{-7}	**5.1×10^{-17}**	0.1031	0.1663	(0.9557)	(2.0×10^{-4})	0.1831
Repressors	ZFP161	K562	HepG2	0.8714	0.9325	(0.2213)	(0.7007)	(0.0281)	(0.5829)	(0.1919)	0.7643	(0.3882)	(0.3739)	(1.0000)	0.7213	(0.5272)
	GFI1	HepG2	K562	(0.0417)	**(0.0369)**	(0.4955)	0.3942	(0.5701)	0.3942	0.7499	0.9001	(0.1556)	(0.1252)	0.9547	(0.7333)	0.1714
	Combined			(0.0490)	(0.0815)	(0.1648)	0.7327	(0.0517)	0.9425	(0.4212)	0.8254	(0.1023)	(0.0819)	0.8689	(0.9036)	0.6570

b

		Enhancer cell line	Expression cell line	High vs. ignored conserv. (160)[a]	High conservation: Scramble (160)[b]	Removal (15)[b]	Max 1-bp decrease (15)[b]	Least 1-bp change (15)[b]	Max 1-bp increase (15)[b]	Random 1-bp change (30)[b]	Ignored conservation: Scramble (160)[b]	Removal (15)[b]	Max 1-bp decrease (15)[b]	Least 1-bp change (15)[b]	Max 1-bp increase (15)[b]	Random 1-bp change (30)[b]
Activators	HNF1	HepG2	K562	0.3728	0.1671	(1.0000)	0.5701	0.5701	0.0995	0.4165	0.8338	(0.3066)	(0.3635)	(0.4955)	(0.0535)	(0.3820)
	HNF4	HepG2	K562	(0.3290)	(0.4116)	(0.7333)	(0.2330)	(0.5321)	0.3942	(0.6143)	(0.8352)	0.2115	0.2330	(0.0995)	(0.7764)	(0.9754)
	FOXA	HepG2	K562	(0.9747)	0.8902	(0.1556)	(0.2560)	(0.1252)	(0.0468)	(0.0020)	(0.1788)	0.9750	(0.7299)	0.6378	0.9750	(0.7847)
	GATA	K562	HepG2	0.0155	0.0440	(0.4603)	0.9547	(0.6092)	(0.3635)	(0.3286)	(0.2652)	(0.5701)	(0.2115)	0.1118	0.7333	0.6884
	NRF2	K562	HepG2	0.0316	**1.1×10^{-12}**	0.2719	0.2209	0.9250	(0.0029)	0.0305	6.5×10^{-8}	0.7221	0.6566	0.9292	(0.2477)	0.0240
	Combined			0.0465	3.3×10^{-6}	(0.5024)	(0.7939)	(0.4237)	(0.1349)	(0.6268)	0.1269	(0.9697)	(0.7058)	0.9697	(0.1108)	0.5944
Repressors	ZFP161	K562	K562	(0.2057)	0.0529	0.1520	0.7532	0.1579	0.1579	0.7366	0.0534	(0.3465)	(0.5337)	(0.8589)	(0.7213)	(0.7164)
	GFI1	HepG2	HepG2	(0.2884)	**(0.5798)**	(0.2115)	(0.6496)	0.9096	(0.1398)	(0.3493)	(0.6216)	(0.1398)	(0.3066)	(0.2330)	(0.1252)	(0.0207)
	Combined			(0.0295)	0.5343	0.6987	(1.0000)	0.3246	0.7731	(0.9332)	0.3398	(0.0926)	(0.2087)	(0.4237)	(0.1919)	(0.0629)

c

		Enhancer cell line	Expression cell line	High vs. ignored conserv. (18)[a]	High conservation: Scramble (18)[b]	Ignored conservation: Scramble (18)[b]
Activators	HNF1	K562	K562	(0.1249)	(0.0347)	(0.5566)
	HNF4	K562	K562	(0.3038)	0.9133	0.5862
	FOXA	K562	K562	(0.4290)	(0.6475)	0.1446
	GATA	HepG2	HepG2	(0.8993)	0.7112	0.9133
	NRF2	HepG2	HepG2	3.5×10^{-4}	0.0018	(0.3958)
	Combined			0.6972	0.4579	0.5828
Repressors	ZFP161	HepG2	K562	(0.6316)	(0.7989)	0.8446
	GFI1	K562	HepG2	(0.7517)	0.9058	0.2145
	Combined			(0.7047)	0.9808	0.3223

$p = 10^{-10}$ $p > 0.05$ $p = 10^{-10}$

higher in modifed lower in modifed

enhancers that showed at least 0.5 expression score (Figure 6-10).

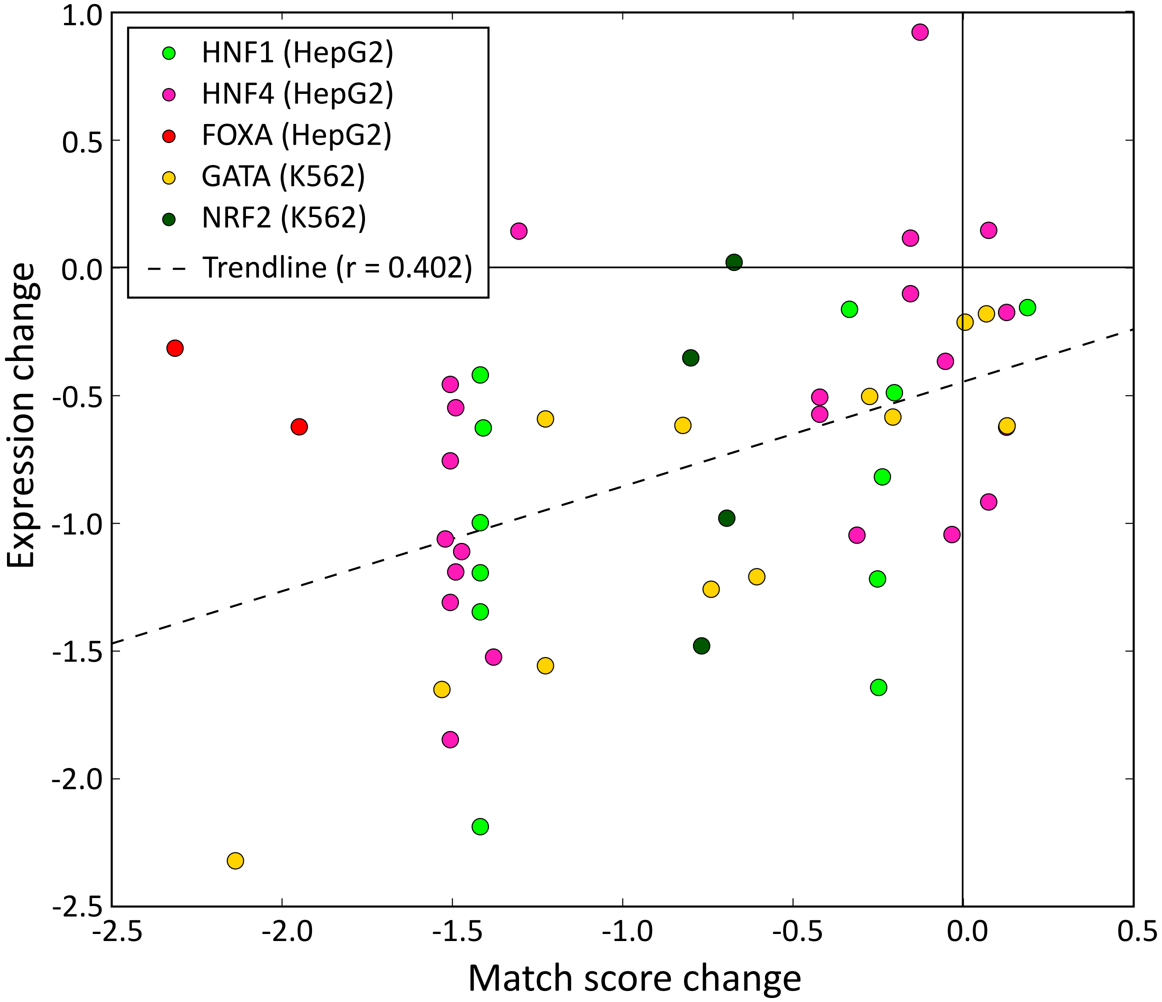


Figure 6-10: Change in expression for random manipulations of conserved instances with wild type expression score of at least 0.5 (52 of the 148 random manipulations). A correlation is seen between the change in match score strength and the corresponding change in expression ($r = 0.402$; random permutation $P = 0.0023$). Match score is normalized for each motif such that 0 is the weakest possible permitted match and 1 is a perfect match (scores can be less than 0 after the manipulations).

We generally did not see an effect for manipulations when the motif and enhancer cell type did not match the cell type where expression was measured (Figure 6-17). A notable exception is NRF2, for which scrambling motif instances found in K562 enhancers also resulted in a significant reduction in expression when measured in the HepG2 cell line ($P_W = 1.1 \times 10^{-12}$). This change in enhancer activity is significantly higher for the instances that are also enhancers in

HepG2 ($P_U = 0.0138$). This suggests that NRF2 is also active in HepG2 and indeed this has been previously reported (Gong and Cederbaum, 2006). Because our initial motif enrichment and expression analysis failed to show this possibility, this suggests that MPRA may be a viable method for performing a *de novo* identification of active factors in a cell line through the systematic evaluation of motif matches in cell line specific enhancers.

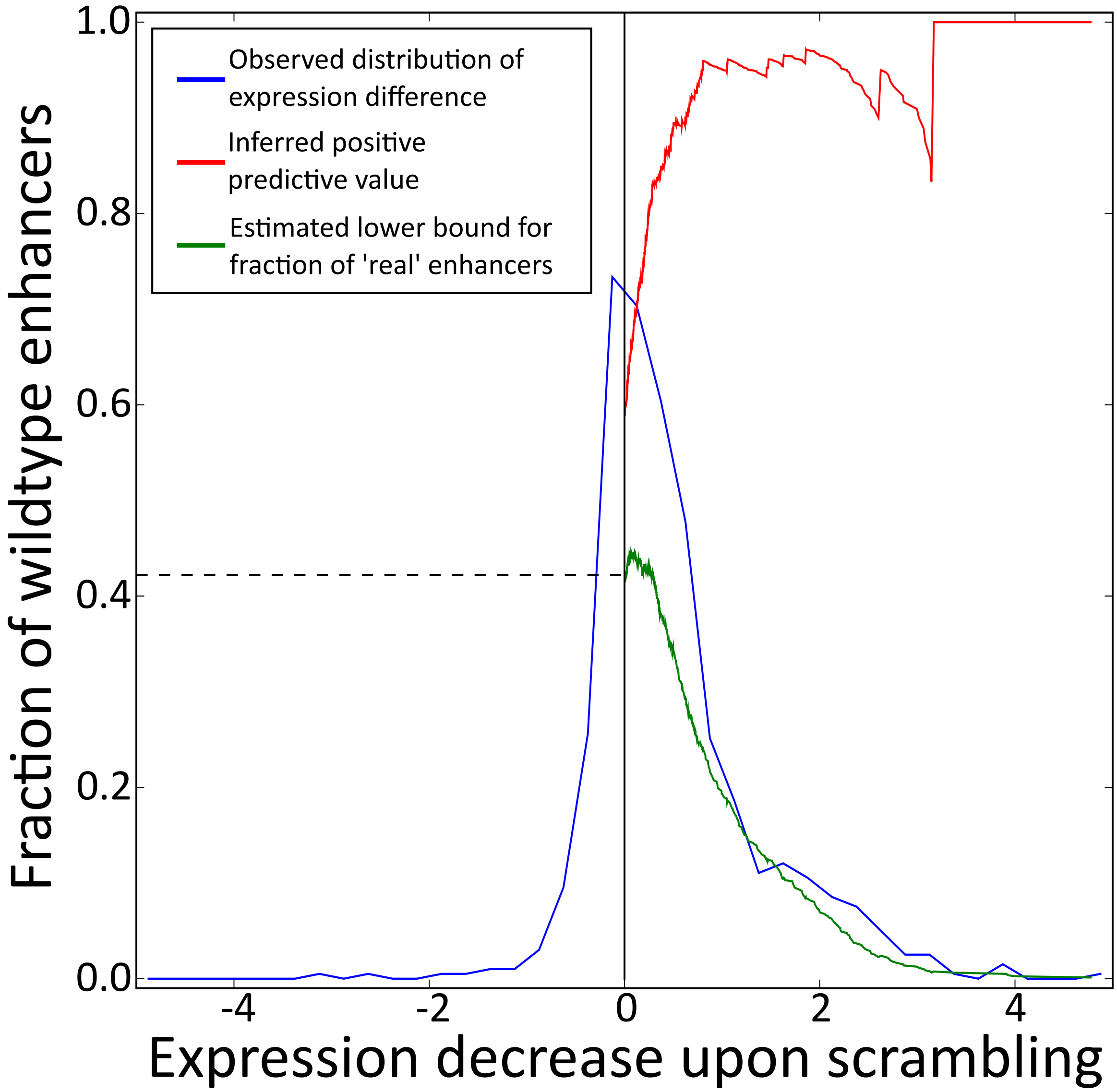


Figure 6-11: We estimate the number of 'real' enhancers by considering the fraction that reduce in expression upon scrambling. Here for comparative motif instances for activators, 71% of tested sequences reduce upon scrambling

6.4.2 Estimating the number of functional tested enhancers

We found that 29% of comparative instances of activators had increased expression in their target cell line upon scrambling. In order to obtain a lower bound on the number of functional enhancers, we conservatively assume that only non-functional enhancers will increase upon scrambling and that non-functional enhancers are equally likely to decrease on scrambling. Consequently, we estimate that $100\% - 2 \times 29 = 42\%$ of the tested enhancers based on comparative motif instances are functional (Figure 6-11). The same analysis for instances ignoring conservation results in an estimate of 23% being real, based on 39% increasing in expression upon scrambling.

We also estimate this fraction of real instances in an independent way using the replicate data and find remarkably similar results. As described above, each sequence is tested with ten distinct tags and two biological replicates, although some are thrown out due to having too few plasmid reads. We compute a one-sided Mann-Whitney p-value for each tested sequence comparing these up to 20 values to the corresponding values for their scramble. We then compute q-values (Benjamini and Hochberg, 1995) from these p-values and estimate a lower bound for the number of functional enhancers at each cutoff by taking the product of the q-value and the number of samples at that q-value. The maximum of these values — 37% for comparative instances and 21% for instances ignoring conservation — is then an independent lower bound estimate of the number of functional enhancers.

6.4.3 Properties of functional enhancers

We examined the enhancers based on instances for activators ignoring conservation to identify features that lead to expression. We first notice that the motif match sequence alone cannot explain this difference (Figure 6-12), consequently the tested context of the sequence must play an important role. We looked at several features of the sequence to identify which could be responsible for this

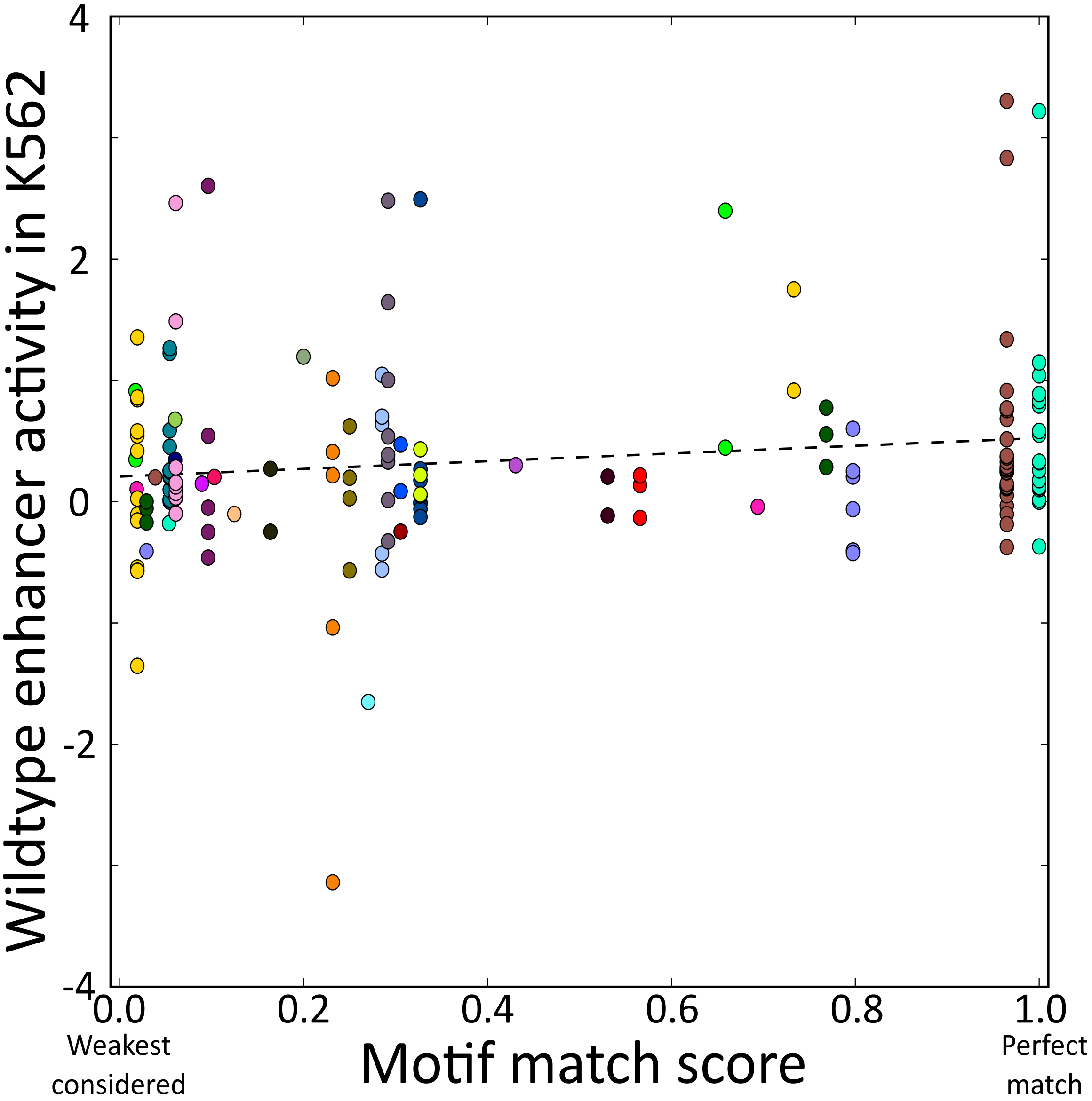


Figure 6-12: The expression level of the wild-type enhancers for NRF2 at different motif strength levels (each color indicates a different match sequence). Variability in expression for each sequence highlights the importance of sequence context.

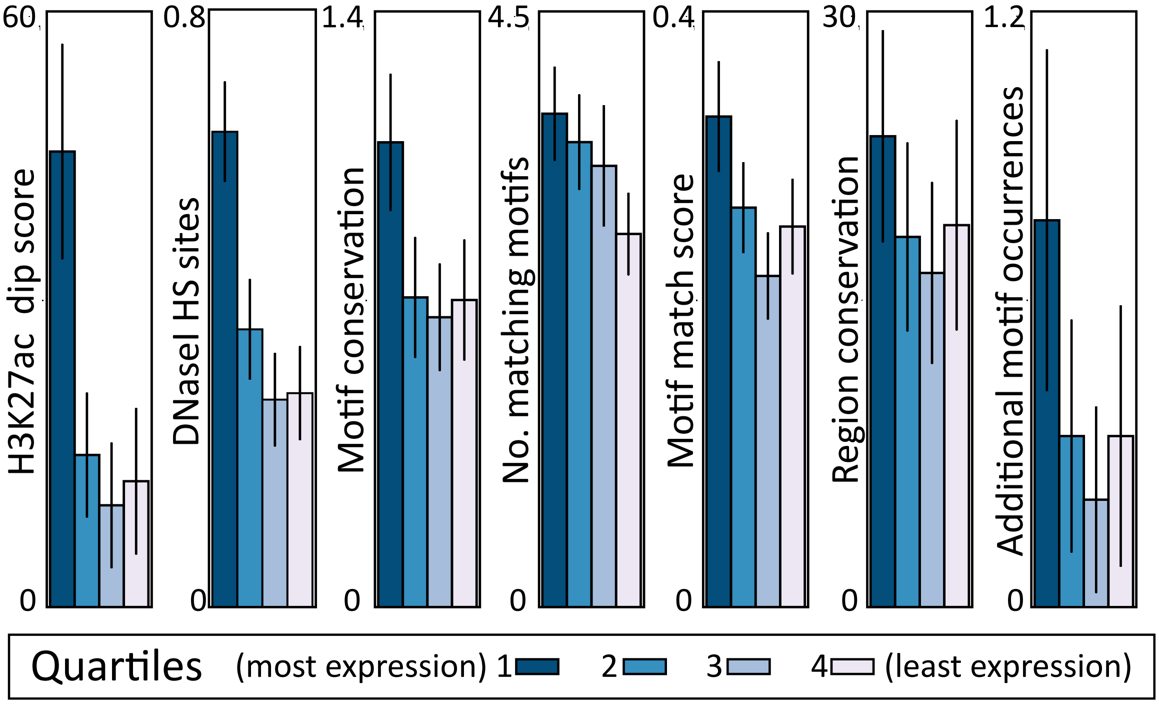


Figure 6-13: Association of the top scoring enhancers with: the average H3K27ac signal value in the matched cell type 200-bp away minus the value centered on the motif (in 25-bp windows); overlap with DNaseI hypersensitivity data matched cell line (Song et al., 2011); the raw BLS score; the number of factors with matching motifs in regions outside the motif match in the tested sequence; the strength of the motif match; the number of bases indicated as conserved by SiPhy-ω 12-mers (Garber et al., 2009); and the number of matches to the tested motif within tested sequence.

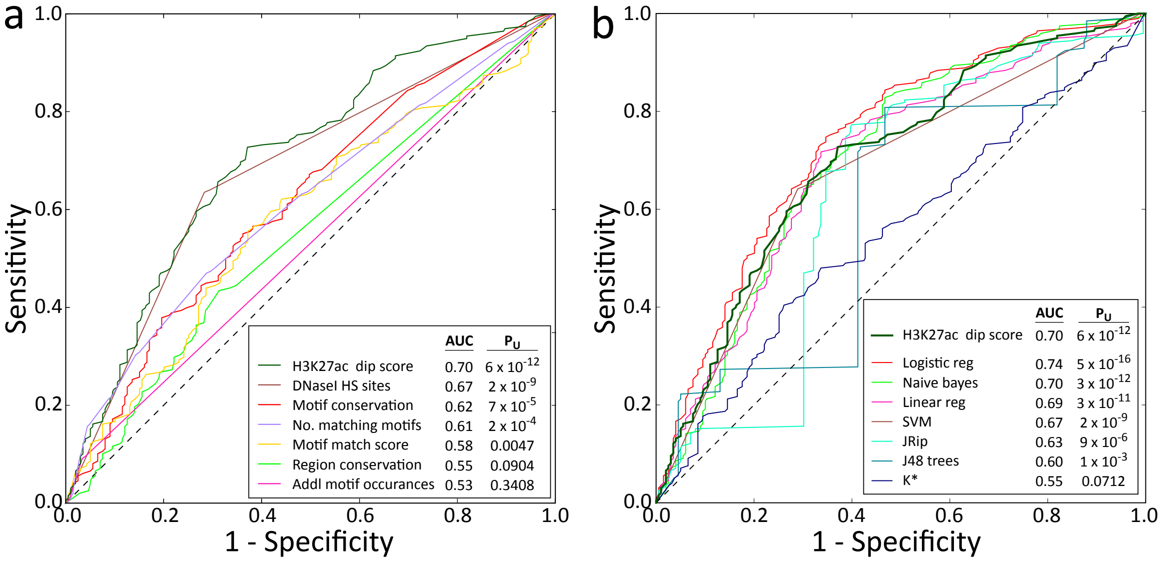


Figure 6-14: (a) Predicted power of each of these features for separating the top and bottom quartile of expression for tested sequences. (b) Combination of the properties shown in (a) using various machine learning techniques and comparison to best individual feature (H3K27ac dip). We employed 7 machine learning techniques implemented by WEKA (v3.6.4; Hall et al., 2009) and scored each of top and bottom 25% of sequences for each activator dataset in a leave-one-out cross-validation framework. Three pairs of sequences overlapped and were placed in the same cross-validation group. We find that the logistic regression algorithm outperforms the best individual feature (AUC 0.74 vs. 0.70).

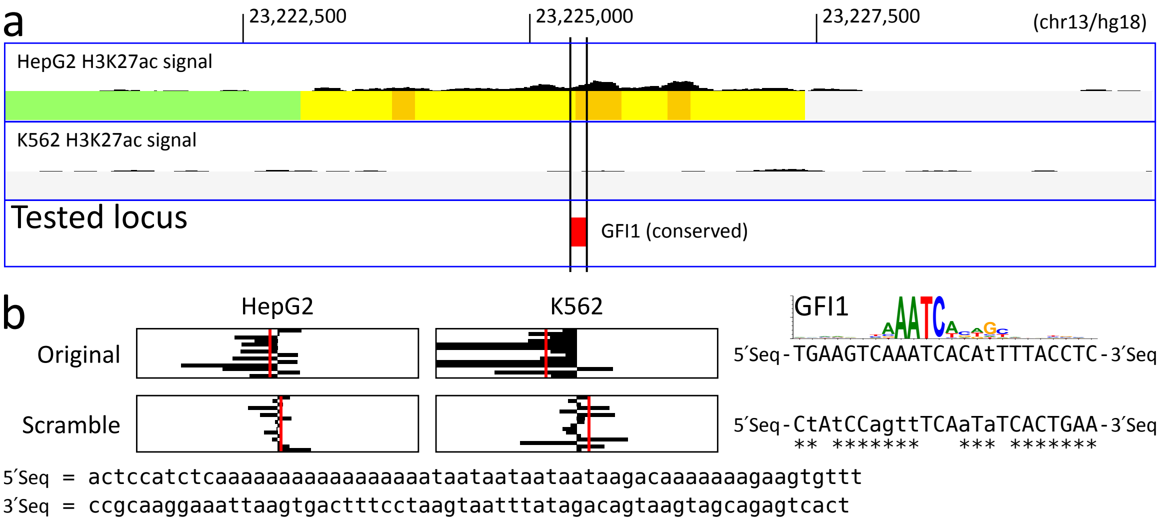


Figure 6-15: Example putative repressor region tested inside HepG2 enhancer. (a) Tested sequence for a repressor, GFI1. A alignment-free conserved GFI1 motif instance inside a HepG2 enhancer is selected. (e) Upon scrambling the binding site, the repression initially seen in K562 (where GFI1 is predicted to be a repressor) is reduced. No change is seen in HepG2 where we do not predict GFI1 to be active.

difference in expression and found that a local reduction in the chromatin signal best differentiated between high and low expressed sequences (Figure 6-13). Moreover, we found that combining these features could result in a modest increase in performance (Figure 6-14), suggesting the non-redundancy of the best feature with the others.

Because MPRA is conducted in plasmids that are not thought to become fully chromatinized and the 145-bp sequences are completely removed from their endogenous sequence context, this suggests a causative role for the TFs in both moving the chromatin and leading to increased expression. However, further *in vivo* studies would be necessary to validate this.

6.4.4 Repressor motifs block enhancer activity

In addition to the five enriched motifs we predicted to be activators, we also consider the case of motif depletion in cell line specific enhancers. We reason that this depletion coupled with expression of the corresponding factor is consistent with a repressor: regions bound by an actively expressed protein would be silenced and

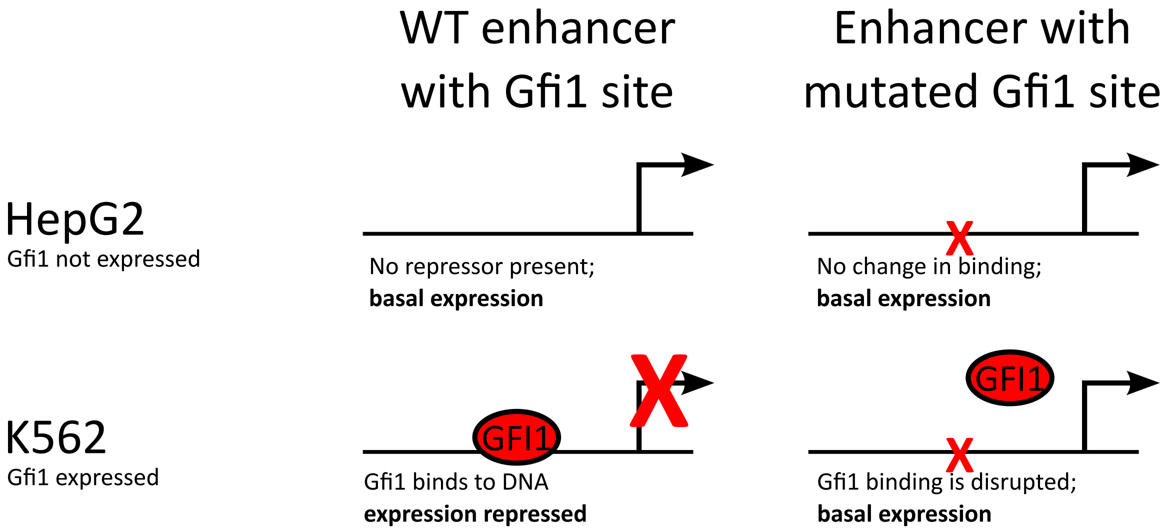


Figure 6-16: Model of action for repressors and the effect of mutating their binding sites. Expression is unchanged in the cell line where enhancers are taken from because the repressor is not present. Conversely, in the unmatched cell line expression is increased upon mutation of the repressor binding site. This model accommodates multiple mechanisms for repressor action including turning off of the moderate SV40 promoter or preventing the binding of unknown activators.

lose enhancer activity. We expect that manipulation of the bases recognized by a repressor could result in aberrant enhancer activity in a cell line where the repressor is expressed. We predict one potential repressor for each cell line: ZFP161 for HepG2 and GFI1 for K562 (Figure 6-3). Because we do not expect functional binding sites of these factors in their corresponding cell line, we instead test sites found in the opposite cell line.

Figure 6-15 shows an example tested putative repressor region, centered on a alignment-free conserved GFI1 motif match in an HepG2 enhancer. We notice a reduction of expression from baseline in the cell line where enhancer activity was not seen (K562). This reduction is then eliminated upon scrambling the binding site for the putative repressor (GFI1). This increase in expression is consistent with our repressor model (Figure 6-16) and the trend we see for comparative GFI1 motif instances in HepG2 enhancers (Wilcoxon $P_W = 0.0369$; Figure 6-9) measured in K562 cells. Conversely, no significant change was seen in HepG2 cells where the enhancers were selected ($P_W = 0.5798$). However, we did not observe any statistically significant difference in expression driven by ZFP161 motif instances in K562 enhancers on the basis of comparative or scrambling in either K562 or HepG2 cells. This may be because we (1) incorrectly predicted ZFP161 as a repressor in HepG2 cells, (2) additional repressive signals that beyond the match to ZFP161 acted redundantly to prevent expression, (3) we were unable to test a sufficient number of sites for significance.

6.5 Conclusion

We have shown that small windows (145 bp) around comparative motif instances are often sufficient to capture cell line specific enhancer activity. Further studies will need to be conducted to predict the specific context cues beyond the motif match that result in expression. Also, the larger context around the tested sequences may have additional effects, either increasing the cell line specific expression or reducing the effect of manipulations due to redundancy. The combinatorial nature of regulation was also not investigated here and is perhaps something that will most benefit from MPRA's capacity to test thousands of sequences in parallel.

6.5.1 Biological contributions

The results of this experiment also lead to a number of biological contributions. We make a clear confirmation of the role that activators play in the establishment of enhancers. Particularly, rather than a view of enhancers as being very highly conserved sequences that tolerate few mutations (Visel et al., 2007), we show that they can and do tolerate neutral ('silent') manipulations. Moreover, we highlight the perhaps under appreciated role that repressors make in restricting inappropriate enhancer activity. Finally, we identify specific features associated with enhancers and quantify their relative importance in the context of the regions we selected — an analysis only made possible by the large number of enhancers we tested here.

Enhancer cell line	Factor	Conser-vation	Original	Scramble	Removal	Max 1-bp decrease	Least 1-bp change	Max 1-bp increase	Random 1-bp change	Total
HepG2	HNF1	high	154[a]	160	15	15	15	15	30	400
		ignored	158[a]	160	15	15	15	15	30	406
	HNF4	high	160	160	15	15	15	15	30	408
		ignored	160	160	15	15	15	15	30	409
	FOXA	high	160	160	15	15	15	15	30	407
		ignored	160	160	14[b]	15	15	15	30	406
	GATA	high	18	18						36
		ignored	18	18						36
	NRF2	high	18	18						36
		ignored	18	18						36
	ZFP161	high	10[d]	10[d]						20
		ignored	18	18						36
	GFI1	high	160	160	15	15	15	15	30	408
		ignored	160	160	15	15	15	15	30	410
K562	HNF1	high	18	18						36
		ignored	18	18						36
	HNF4	high	18	18						36
		ignored	18	18						36
	FOXA	high	18	18						36
		ignored	17[a]	17[a]						34
	GATA	high	160	160	15	15	15	15	30	408
		ignored	160	160	15	15	15	15	30	408
	NRF2	high	159[b]	159[b]	14[c]	14[c]	14[c]	14[c]	28[c]	400
		ignored	160	160	12[c]	12[c]	12[c]	12[c]	24[c]	392
	ZFP161	high	51[d]	51[d]	15	15	15	15	30	191
		ignored	105[d]	105[d]	15	15	15	15	30	299
	GFI1	high	18	18						36
		ignored	18	18						36
Total			2104	2112	203	204	204	204	412	5418

Table 6-2: Precise number of tested sequences. Number fewer than indicated in Table 6-1 due to [a]identical sequences being found at different locations or due to matches on opposite strands, [b]the creation of a restriction site at boundary of tested region, [c]the motif instance being the best possible match to the desired motif, and consequently excluded from all non-scramble manipulations, or [d]having too few matches in the desired enhancer type. Totals indicate the number of distinct tested sequences and thus differ from sum of the corresponding rows or columns due to reuse of sequences across modifications (e.g., a random mutation matching one of the directed 1-bp modifications) or reuse of a position (e.g., a comparative instance tested even when conservation was not taken into account). These identical sequences are tested only once.

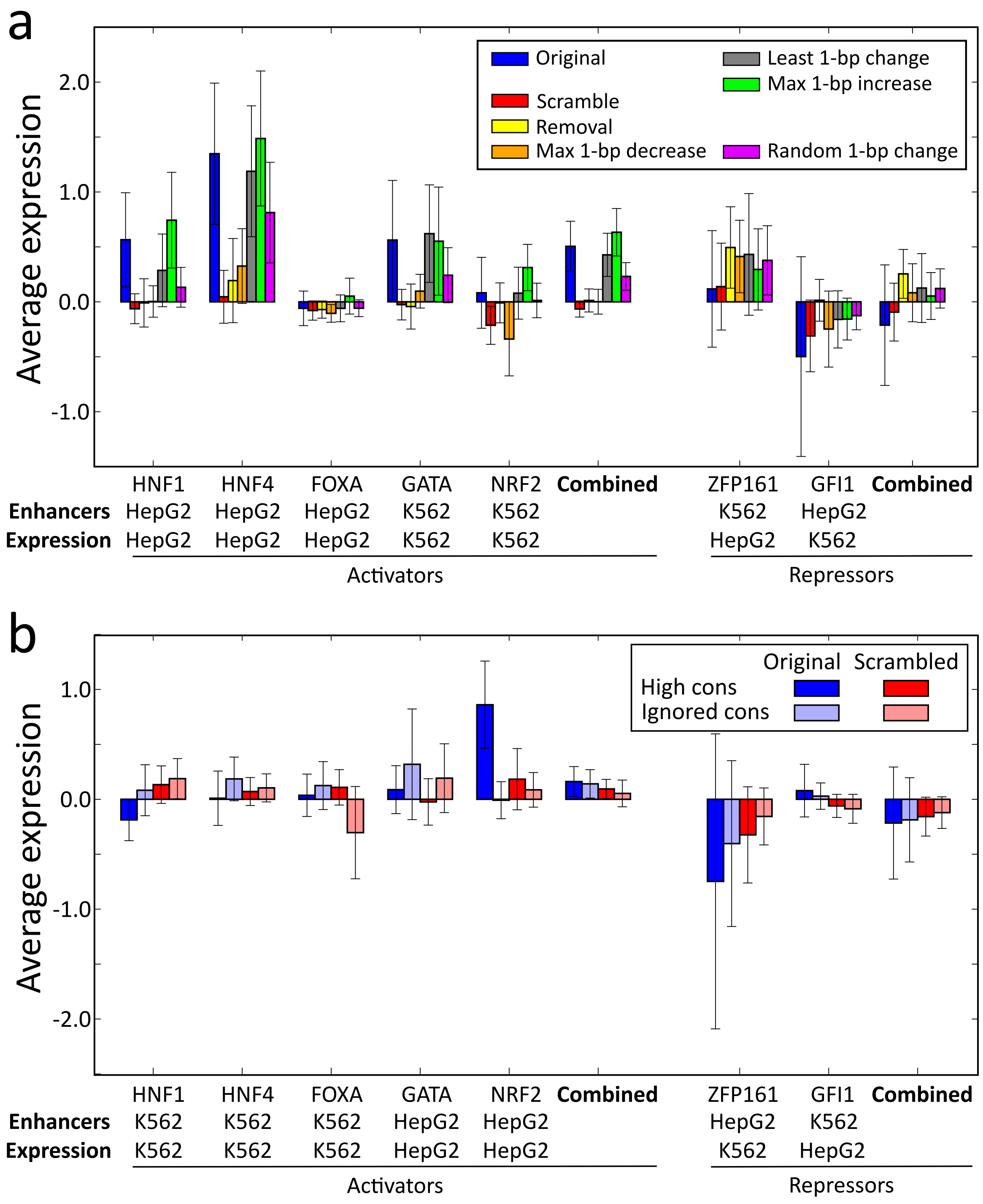


Figure 6-17: (a) Bar plots showing differences amongst locations tested for additional modifications. Because these are restricted to only these regions, the original and scramble value differs from those in Figure 6-8 (b) As with Figure 6-8, except on the opposite cell types for both enhancer selection and cell type (using the 18 control enhancers tested per factor). Lack of significant difference is seen for every factor except NRF2.

Chapter 7

Conclusion

7.1 Summary of results

This thesis presented computational approaches for integrating diverse datasets with the goal of understanding the regulation of animal genomes.

We began with a novel, robust empirically driven algorithm for predicting functional motif instances using comparative genomics (Chapter 2). Our approach is able to predict instances in complex animal genomes where limited sequencing, alignment errors, and evolutionary turn over is commonplace with a score correcting for motif composition and complexity. We found that the resulting motif instances were competitive with ChIP-chip/seq in predicting genes where we expect a regulatory relationship, but can be produced without experiments or antibodies. The core of our algorithm is a matching program which we engineered to be fast enough to produce instances using dozens of large, mammalian genomes. These motif instances build a network relationship between genes and are used extensively throughout the thesis, highlighting their practicality.

In the next two chapters we explored two ways motifs could computationally be predicted. In Chapter 3 we utilize motif discovery in ChIP-seq datasets to better understand the relationship between transcription factors. We are able to recover the known specificity for the majority of TFs and, moreover, we find additional enriched motifs for most datasets. In turn, these additional motifs predict key regulator relationships, many of which we validate as either synergistic or antagonistic through an extensive literature search. We also find several motifs that had not been previously described.

We predict motifs in Chapter 4 through the *de novo* computational discovery of miRNAs and the characterization of their $5'$ mature cleavage. Beyond the unmatched performance of these algorithms, we find a remarkable connection between our features and the underlying biology: miRNAs that were easy for our algorithm to find were also expressed at higher levels by the organism and when we had difficultly predicting the correct processing position, so did the cell.

Lastly, we sought how we could go beyond our static motif instances and understand the fundamentally dynamic nature of development. In Chapter 5 we predicted regulators, both activators and repressors, using specific signatures of expression and motif enrichment in chromatin states for nine human cell lines. We first supported these predicted regulators through an analysis of known biology for the cell lines we considered. We then systematically tested our predictions using a hypothesis driven approach centered around engineered sequences based on our motif instances (Chapter 6). We found that we were able to predict functional enhancers using our comparative motif instances at rates comparable to *in vivo* enhancers identified using much longer, highly conserved sequences. Again, our predictions and their subsequent validation allowed us to learn a number of biological lessons, including the features associated with functional enhancers and the behavior of activators versus repressors.

7.2 Future work

The work presented in this thesis can be extended in a variety of ways, particularly as technology and available data increases.

As we argued in Chapter 2, mammalian comparative motif discovery does not greatly benefit from many additional species, and we reason that the same is true for other clades (e.g., yeasts and flies). Conversely, we found that for both motif instances (Figure 2-10) and miRNAs (Figure 4-7) additional species found in appropriate places in the phylogeny would continue to increase discovery power. Consequently, we expect that as time passes and the number of available genomes goes from dozens to thousands the utility of these algorithms and their predictions will increase.

In Chapter 3 we saw significant differences in the quality of known motifs for transcription factors. A currently unexplored area in comparative regulatory genomics is using conservation not for *de novo* motif discovery, but rather for refining known motifs. There are several challenges for this type of approach, including how to integrate phylogenetic data, deal with problematic alignments, and assess performance. However, comparative and experimental data may complement each other here as we have shown they do elsewhere in this thesis.

Moreover, while we used our comparative motif instances to understand chromatin dynamics and TF binding, our motif instances may also be useful for directly understanding combinatorial binding of transcription factors. For example, we could examine the frequency of conserved motif instances occurring in the vicinity of each other, or the relative orientations — two analysis that cannot be done using experimental binding data alone.

To increase the impact of our motif instances, we need to be able to link them to genes. For instances very close to genes (e.g., within 2kb) their linking is relatively unambiguous, something we take advantage of in Chapter 2. However, more than 90% of the human genome is more than 2kb away from a transcription start site and known enhancers have been found hundreds of kilobases away from their target TSS (Nobrega et al., 2003; Lettice et al., 2003). New technologies such as 3C (Dekker et al., 2002), Hi-C (Lieberman-Aiden et al., 2009), and ChIA-PET (Fullwood et al., 2009) promise to identify links between genes and distant regulatory loci. This may allow for a dramatically higher sensitivity in predicting which motifs regulator which genes.

We showed that MPRA is often able to recapitulate *in vivo* enhancer activity outside of the original context using a 145-bp enhancer region centered on a alignment-free conserved motif instance. However, it remains to be seen how the motif and the remaining $\sim$135-bp interact. Rather than a stochastic mutational approach on a small number of loci, a hypothesis driven design focusing on motif matches may be fruitful. Moreover, we believe MPRA could be useful in predicting specific regulators that are active in a cell type, overcoming problems with expression analysis stemming from protein-level modification.

While advances in experimental techniques seem inevitable, they are equally unpredictable. Consequently, while the specific means through which insights in regulatory genomics will occur is unknown, the continued increase in knowledge about this important aspect of molecular biology seems assured. This promise ensures that the future of computational biology will be as exciting as recent years have been.

Bibliography

Abrams, E. W. and Andrew, D. J., 2005. CrebA regulates secretory activity in the *Drosophila* salivary gland and epidermis. *Development*, **132**(12):2743–2758.

Adams, B., Dörfler, P., Aguzzi, A., Kozmik, Z., Urbánek, P., Maurer-Fogy, I., and Busslinger, M., 1992. Pax-5 encodes the transcription factor BSAP and is expressed in b lymphocytes, the developing CNS, and adult testis. *Genes & Development*, **6**(9):1589 –1607.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.*, 2000. The genome sequence of *Drosophila melanogaster*. *Science*, **287**(5461):2185 –2195.

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., and Moor, B. D., 2003. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Research*, **31**(6):1753 –1764.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P., 2002. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 4 edition.

Alterovitz, G., Benson, R., and Ramoni, M. F., 2009. *Automation in proteomics and genomics: an engineering case-based approach*. John Wiley and Sons.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3):403–410.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., *et al.*, 2009. Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**(5935):1720–1723.

Baeuerle, P. A. and Henkel, T., 1994. Function and activation of NF-kappa b in the immune system. *Annual Review of Immunology*, **12**:141–179.

Bailey, T. L. and Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, **2**:28–36.

Bailey, T. L. and Gribskov, M., 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**(1):48–54.

Bairoch, A., 2004. The universal protein resource (UniProt). *Nucleic Acids Research*, **33**(Database issue):D154–D159.

Baliga, N. S., 2001. Promoter analysis by saturation mutagenesis. *Biological Procedures Online*, **3**:64–69.

Bar-Joseph, Z., Gifford, D. K., and Jaakkola, T. S., 2001. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, **17**:S22–S29.

Barski, A., Cuddapah, S., Cui, K., Roh, T., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K., 2007. High-Resolution profiling of histone methylations in the human genome. *Cell*, **129**(4):823–837.

Bartel, D., 2009. MicroRNAs: target recognition and regulatory functions. *Cell*, **136**(2):215–233.

Bartel, D. P., 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**(2):281–297.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S., 2002. ARACHNE: a Whole-Genome shotgun assembler. *Genome Research*, **12**(1):177 –189.

Beal, K., Kheradpour, P., Wilder, S., Kundaje, A., Dunham, I., Kellis, M., Birney, E., and Herrero, J., 2011. Modulation of the transcription factor affinity in the human genome. Submitted, GRCP042.

Bender, W., 2008. MicroRNAs in the *Drosophila* bithorax complex. *Genes & Development*, **22**(1):14 –19.

Benjamin J., B., 2006. Alternative splicing: New insights from global analyses. *Cell*, **126**(1):37–47.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**(1):289–300.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., *et al.*, 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*, **37**(7):766–770.

Berezikov, E., Cuppen, E., and Plasterk, R. H. A., 2006. Approaches to microRNA discovery. *Nature Genetics*, **38 Suppl**:S2–7.

Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Peña-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T., *et al.*, 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**(7):1266–1276.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L., 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, **24**(11):1429–1435.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(2):757–762.

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

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., *et al.*, 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**(7146):799–816.

Bischof, L. J., Kagawa, N., Moskow, J. J., Takahashi, Y., Iwamatsu, A., Buchberg, A. M., and Waterman, M. R., 1998. Members of the meis1 and pbx homeodomain protein families cooperatively bind a cAMP-responsive sequence (CRS1) from BovineCYP17. *Journal of Biological Chemistry*, **273**(14):7941 –7948.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.*, 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, **14**(4):708 –715.

Blanchette, M. and Tompa, M., 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research*, **12**(5):739–48.

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M., and Snyder, M., 2007. Divergence of transcription factor binding sites across related yeast species. *Science*, **317**(5839):815 –819.

Breiman, L., 2001. Random forests. *Machine Learning*, **45**(1):5–32.

Buck, M. J. and Lieb, J. D., 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**(3):349–360.

Cahir McFarland, E. D., Izumi, K. M., and Mosialos, G., 1999. Epstein-barr virus transformation: involvement of latent membrane protein 1-mediated activation of NF-kappaB. *Oncogene*, **18**(49):6959–6964.

Caretti, G., Salsi, V., Vecchi, C., Imbriano, C., and Mantovani, R., 2003. Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters. *Journal of Biological Chemistry*, **278**(33):30435 –30440.

Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., *et al.*, 2009. Unlocking the secrets of the genome. *Nature*, **459**(7249):927–930.

Chan, H. M. and La Thangue, N. B., 2001. p300/CBP proteins: HATs for transcriptional bridges and scaffolds. *Journal of Cell Science*, **114**(13):2363 –2373.

Che, D., Jensen, S., Cai, L., and Liu, J. S., 2005. BEST: binding-site estimation suite of tools. *Bioinformatics*, **21**(12):2909–2911.

Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., and Childs, G., 1999. Making and reading microarrays. *Nature Genetics*, **21**(1 Suppl):15–19.

Choi, S., Cho, Y., Kim, H., and Park, J., 2007. ROS mediate the hypoxic repression of the hepcidin gene by inhibiting C/EBPalpha and STAT-3. *Biochemical and Biophysical Research Communications*, **356**(1):312–317.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M., 2003. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, **301**(5629):71 –76.

Consortium, T. E. P., 2011a. Initial analysis of the encyclopedia of DNA elements in the human genome. Submitted, NCP000.

Consortium, T. E. P., 2011b. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, **9**(4):e1001046.

Consortium, T. m., Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., *et al.*, 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, **330**(6012):1787 –1797.

Corcoran, D. L., Pandit, K. V., Gordon, B., Bhattacharjee, A., Kaminski, N., and Benos, P. V., 2009. Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS ONE*, **4**(4):e5279.

Corcoran, L. M., Karvelas, M., Nossal, G. J., Ye, Z. S., Jacks, T., and Baltimore, D., 1993. Oct-2, although not required for early b-cell development, is critical for later b-cell maturation and for postnatal survival. *Genes & Development*, **7**(4):570 –582.

Core, L. J. and Lis, J. T., 2008. Transcription regulation through Promoter-Proximal pausing of RNA polymerase II. *Science*, **319**(5871):1791 –1792.

Costa, R. H., Kalinichenko, V. V., Holterman, A. L., and Wang, X., 2003. Transcription factors in liver development, differentiation, and regeneration. *Hepatology*, **38**(6):1331–1347.

Crooks, G. E., Hon, G., Chandonia, J., and Brenner, S. E., 2004. WebLogo: a sequence logo generator. *Genome Research*, **14**(6):1188–1190.

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T., Cui, K., and Zhao, K., 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, **19**(1):24–32.

Davidson, E. H., McClay, D. R., and Hood, L., 2003. Regulatory gene networks and the properties of the developmental process. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(4):1475–1480.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N., 2002. Capturing chromosome conformation. *Science*, **295**(5558):1306 –1311.

Deng, C., 2003. Roles of BRCA1 in DNA damage repair: a link between development and cancer. *Human Molecular Genetics*, **12**(90001):113R–123.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M., 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, **14**(4):457–460.

Down, T. A., Bergman, C. M., Su, J., and Hubbard, T. J. P., 2007. Large-Scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Computational Biology*, **3**(1):e7.

Drosophila 12 Genomes Consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**(7167):203–218.

Dudek, H., Tantravahi, R. V., Rao, V. N., Reddy, E. S., and Reddy, E. P., 1992. Myb and ets proteins cooperate in transcriptional activation of the mim-1 promoter. *Proceedings of the National Academy of Sciences*, **89**(4):1291 –1295.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M., 1999. Expression profiling using cDNA microarrays. *Nature Genetics*, **21**(1 Suppl):10–14.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G., 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Ernst, J. and Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, **28**(8):817–825.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.*, 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345):43–49.

Ernst, J., Vainas, O., Harbison, C. T., Simon, I., and Bar-Joseph, Z., 2007. Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, **3**.

Ettwiller, L., Paten, B., Ramialison, M., Birney, E., and Wittbrodt, J., 2007. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature Methods*, **4**(7):563–565.

Ettwiller, L., Paten, B., Souren, M., Loosli, F., Wittbrodt, J., and Birney, E., 2005. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biology*, **6**(12):R104.

Farh, K. K., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B., and Bartel, D. P., 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, **310**(5755):1817–1821.

Farnham, P. J., 2009. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, **10**(9):605–616.

Fitzsimmons, D., Hodsdon, W., Wheat, W., Maira, S. M., Wasylyk, B., and Hagman, J., 1996. Pax-5 (BSAP) recruits ets proto-oncogene family proteins to form functional ternary complexes on a b-cell-specific promoter. *Genes & Development*, **10**(17):2198 –2211.

Fleiss, J. L., Levin, B., Paik, M. C., and Fleiss, J., 2003. *Statistical Methods for Rates & Proportions*. Wiley-Interscience, 3rd edition.

Foat, B. C., Morozov, A. V., and Bussemaker, H. J., 2006. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**(14):e141 –e149.

Forte, E. and Luftig, M. A., 2009. MDM2-Dependent inhibition of p53 is required for Epstein-Barr virus B-Cell growth transformation and Infected-Cell survival. *Journal of Virology*, **83**(6):2491 –2499.

Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N., 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology*, **26**(4):407–415.

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, **19**(1):92 –105.

Frietze, S., Lan, X., Jin, V. X., and Farnham, P. J., 2010. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *Journal of Biological Chemistry*, **285**(2):1393 –1403.

Frith, M. C., Fu, Y., Yu, L., Chen, J., Hansen, U., and Weng, Z., 2004. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Research*, **32**(4):1372 –1381.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., *et al.*, 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**(7269):58–64.

Furusawa, T. and Cherukuri, S., 2010. Developmental function of HMGN proteins. *Biochimica Et Biophysica Acta*, **1799**(1-2):69–73.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X., 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**(12):i54 –i62.

Garbis, S., Lubec, G., and Fountoulakis, M., 2005. Limitations of current proteomics technologies. *Journal of Chromatography A*, **1077**(1):1–18.

Gong, P. and Cederbaum, A. I., 2006. Transcription factor nrf2 protects HepG2 cells against CYP2E1 plus arachidonic acid-dependent toxicity. *Journal of Biological Chemistry*, **281**(21):14573 –14579.

Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P. M., *et al.*, 2010. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. *Nature Methods*, **7**(1):47–49.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J., 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, **36**(Database issue):D154–158.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., and Hofacker, I. L., 2008. The vienna RNA websuite. *Nucleic Acids Research*, **36**(Web Server):W70–W74.

Guo, Y., Papachristoudis, G., Altshuler, R. C., Gerber, G. K., Jaakkola, T. S., Gifford, D. K., and Mahony, S., 2010. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, **26**(24):3028 –3034.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., 2009. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, **11**(1):10–18.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., Lagarde, J., Gilbert, J. G. R., Storey, R., Swarbreck, D., *et al.*, 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, **7 Suppl 1**:S4.1–9.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B., Barrell, D., Zadissa, A., Searle, S., *et al.*, 2011. GENCODE: The reference human genome annotation for the ENCODE project. Submitted, GRCP001.

Haverty, P. M., Hansen, U., and Weng, Z., 2004. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Research*, **32**(1):179 –188.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., *et al.*, 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**(7243):108–112.

Ho Sui, S. J., Mortimer, J. R., Arenillas, D. J., Brumm, J., Walsh, C. J., Kennedy, B. P., and Wasserman, W. W., 2005. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Research*, **33**(10):3154 –3164.

Hock, H. and Orkin, S. H., 2006. Zinc-finger transcription factor Gfi-1: versatile regulator of lymphocytes, neutrophils and hematopoietic stem cells. *Current Opinion in Hematology*, **13**(1):1–6.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chemical Monthly*, **125**(2):167–188.

Huang, X., Wang, J., Aluru, S., Yang, S., and Hillier, L., 2003. PCAP: a Whole-Genome assembly program. *Genome Research*, **13**(9):2164 –2170.

Huang, Y., Myers, S. J., and Dingledine, R., 1999. Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nature Neuroscience*, **2**(10):867–872.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M., 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *Journal of Molecular Biology*, **296**(5):1205–1214.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2):249–264.

Ito, T., Yamauchi, M., Nishina, M., Yamamichi, N., Mizutani, T., Ui, M., Murakami, M., and Iba, H., 2001. Identification of SWI.SNF complex subunit BAF60a as a determinant of the transactivation potential of Fos/Jun dimers. *The Journal of Biological Chemistry*, **276**(4):2852–2857.

Ivan, A., Halfon, M., and Sinha, S., 2008. Computational discovery of cis-regulatory modules in *Drosophila* without prior knowledge of motifs. *Genome Biology*, **9**(1):R22.

Ivanov, V. N., Bhoumik, A., Krasilnikov, M., Raz, R., Owen-Schaub, L. B., Levy, D., Horvath, C. M., and Ronai, Z., 2001. Cooperation between STAT3 and c-Jun suppresses fas transcription. *Molecular Cell*, **7**(3):517–528.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O., 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**(6819):533–538.

Jelinic, P., Stehle, J., and Shaw, P., 2006. The Testis-Specific factor CTCFL cooperates with the protein methyltransferase PRMT7 in h19 imprinting control region methylation. *PLoS Biology*, **4**(11):e355.

Joachims, T., 1999. *Making large-scale support vector machine learning practical*. MIT Press, Cambridge, MA, USA.

Johnson, C. A. and Turner, B. M., 1999. Histone deacetylases: complex transducers of nuclear signals. *Seminars in Cell & Developmental Biology*, **10**(2):179–188.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B., 2007. Genome-Wide mapping of in vivo Protein-DNA interactions. *Science*, **316**(5830):1497–1502.

Jones, N. C. and Pevzner, P. A., 2004. *An Introduction to Bioinformatics Algorithms*. The MIT Press, 1 edition.

Kaiser, J., 2008. A plan to capture human diversity in 1000 genomes. *Science*, **319**(5862):395.

Kappel, A., Schlaeger, T. M., Flamme, I., Orkin, S. H., Risau, W., and Breier, G., 2000. Role of SCL/Tal-1, GATA, and ets transcription factor binding sites for the regulation of flk-1 expression during murine vascular development. *Blood*, **96**(9):3078–3085.

Karin, M., Liu, Z.-g., and Zandi, E., 1997. AP-1 function and regulation. *Current Opinion in Cell Biology*, **9**(2):240–246.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., *et al.*, 2010. Variation in transcription factor binding among humans. *Science*, **328**(5975):232–235.

Kawana, M., Lee, M. E., Quertermous, E. E., and Quertermous, T., 1995. Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. *Molecular and Cellular Biology*, **15**(8):4225–4231.

Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E. S., 2004. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, **11**(2-3):319–355.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S., 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**(6937):241–254.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D., 2002. The human genome browser at UCSC. *Genome Research*, **12**(6):996 –1006.

Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kahari, A., *et al.*, 2009. Ensembl genomes: Extending ensembl across the taxonomic space. *Nucleic Acids Research*, **38**:D563–D569.

Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., *et al.*, 2011. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*, **471**(7339):480–485.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J., 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, **26**(12):1351–1359.

Kheradpour, P., Stark, A., Roy, S., and Kellis, M., 2007. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Research*, **17**(12):1919–1931.

Kim, H. D., Shay, T., O'Shea, E. K., and Regev, A., 2009. Transcriptional regulatory circuits: Predicting numbers from alphabets. *Science*, **325**(5939):429 –432.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H., 2008. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**(6):1049–1061.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V., and Ren, B., 2007. Analysis of the vertebrate insulator protein CTCF-Binding sites in the human genome. *Cell*, **128**(6):1231–1245.

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

Kozomara, A. and Griffiths-Jones, S., 2010. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, **39**(Database):D152–D157.

Kundaje, A., Hariharan, M., Hussami, N., Boyle, A. P., Ouyang, Z., Cheng, Y., Landt, S., Sidow, A., Batzoglou, S., and Snyder, M., *et al.*, 2011a. Context-specific functional associations of transcription factors and their effect on gene expression. Submitted, GRCP018.

Kundaje, A., Jung, Y. L., Kharchenko, P. V., Wold, B., Sidow, A., Batzoglou, S., and Park, P. J., 2011b. Assessment of ChIP-seq data quality using strand cross-correlation analysis. Submitted, GBCP019.

Kundaje, A., Li, Q., Brown, B., Rozowsky, J., Harmanci, A., Wilder, S., Batzoglou, S., Dunham, I., Gerstein, M., Birney, E., *et al.*, 2011c. Reproducibility measures for automatic threshold selection and quality control in chip-seq datasets. Submitted, GRCP020.

Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M., 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biology*, **4**(7):R42.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.*, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921.

Landt, S. G., Marinov, G. K., Kheradpour, P., Kundaje, A., Pauli, F., Batzoglou, S., Bernstein, B., Bickel, P., Brown, B., Cayting, P., *et al.*, 2011. Chip-seq guidelines and practices used by the ENCODE and modENCODE consortia. Submitted, GRCP068.

Latchman, D., 2010. *Gene Control*. Garland Science.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P., 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**(5543):858–862.

Law, J. C., Ritke, M. K., Yalowich, J. C., Leder, G. H., and Ferrell, R. E., 1993. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. *Leukemia Research*, **17**(12):1045–1050.

Lee, C. S., Friedman, J. R., Fulmer, J. T., and Kaestner, K. H., 2005. The initiation of liver development is dependent on foxa transcription factors. *Nature*, **435**(7044):944–947.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E., *et al.*, 2003. A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, **12**(14):1725 –1735.

Lewis, B. P., Burge, C. B., and Bartel, D. P., 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**(1):15–20.

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B., 2003. Prediction of mammalian MicroRNA targets. *Cell*, **115**(7):787–798.

Li, X.-y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C. L. L., *et al.*, 2008. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biology*, **6**(2):e27.

Li-Weber, M., Davydov, I., Krafft, H., and Krammer, P., 1994. The role of NF-Y and IRF-2 in the regulation of human IL-4 gene expression. *The Journal of Immunology*, **153**(9):4122 –4133.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.*, 2009. Comprehensive mapping of Long-Range interactions reveals folding principles of the human genome. *Science*, **326**(5950):289 –293.

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., and Bartel, D. P., 2003a. Vertebrate MicroRNA genes. *Science*, **299**(5612):1540.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., and Bartel, D. P., 2003b. The microRNAs of *Caenorhabditis elegans*. *Genes & Development*, **17**(8):991 –1008.

Lin, C., Vega, V. B., Thomsen, J. S., Zhang, T., Kong, S. L., Xie, M., Chiu, K. P., Lipovich, L., Barnett, D. H., Stossi, F., *et al.*, 2007a. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genetics*, **3**(6):e87.

Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C., Park, S., Wan, K. H., Schroeder, A. J., Gramates, L. S., St. Pierre, S. E., *et al.*, 2007b. Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Research*, **17**(12):000.

Lin, M. F., Deoras, A. N., Rasmussen, M. D., and Kellis, M., 2008. Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes. *PLoS Computational Biology*, **4**(4):e1000067.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., *et al.*, 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**(7370):476–482.

Liu, R., McEachin, R. C., and States, D. J., 2003. Computationally identifying novel NF-kappaB-Regulated immune genes in the human genome. *Genome Research*, **13**(4):654 –661.

Liu, X. S., Brutlag, D. L., and Liu, J. S., 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, **20**(8):835–9.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H., and Matsudaira, P., 2007. *Molecular Cell Biology*. W. H. Freeman, 6th edition.

Loh, Y., Wu, Q., Chew, J., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., *et al.*, 2006. The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, **38**(4):431–440.

Looijenga, L. H. J., Stoop, H., de Leeuw, H. P. J. C., de Gouveia Brazao, C. A., Gillis, A. J. M., van Roozendaal, K. E. P., van Zoelen, E. J. J., Weber, R. F. A., Wolffenbuttel, K. P., van Dekken, H., *et al.*, 2003. POU5F1 (OCT3/4) identifies cells with pluripotent potential in human germ cell tumors. *Cancer Research*, **63**(9):2244–2250.

MacArthur, S., Li, X., Li, J., Brown, J., Chu, H. C., Zeng, L., Grondona, B., Hechmer, A., Simirenko, L., Keranen, S., *et al.*, 2009. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biology*, **10**(7):R80.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T., 2007. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, **35**(Database):D26–D31.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y., Chen, Z., *et al.*, 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057):376–380.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., *et al.*, 2003. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, **31**(1):374–378.

Mazars, R., Gonzalez-de-Peredo, A., Cayrol, C., Lavigne, A., Vogel, J. L., Ortega, N., Lacroix, C., Gautier, V., Huet, G., Ray, A., *et al.*, 2010. The THAP-zinc finger protein THAP1 associates with coactivator HCF-1 and O-GlcNAc transferase: a link between DYT6 and DYT3 dystonias. *The Journal of Biological Chemistry*, **285**(18):13364–13371.

McKay, M. J., Troelstra, C., van der, P., Kanaar, R., Smit, B., Hagemeijer, A., Bootsma, D., and Hoeijmakers, J. H. J., 1996. Sequence conservation of therad21 schizosaccharomyces pombeDNA Double-Strand break repair gene in human and mouse. *Genomics*, **36**(2):305–315.

McLeay, R. C. and Bailey, T. L., 2010. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**(1):165.

Melnikov, A., Murugan, A., Zhang, X., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Jr., Kinney, J. B., *et al.*, 2012. Rapid dissection and model-based optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, **in press**.

Moorman, C., Sun, L. V., Wang, J., de Wit, E., Talhout, W., Ward, L. D., Greil, F., Lu, X., White, K. P., Bussemaker, H. J., *et al.*, 2006. Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(32):12027–12032.

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

Moses, A., Chiang, D., Pollard, D., Iyer, V., and Eisen, M., 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, **5**(12):R98.

Mostoslavsky, R., Chua, K. F., Lombard, D. B., Pang, W. W., Fischer, M. R., Gellon, L., Liu, P., Mostoslavsky, G., Franco, S., Murphy, M. M., *et al.*, 2006. Genomic instability and aging-like phenotype in the absence of mammalian SIRT6. *Cell*, **124**(2):315–329.

Mouthon, M. A., Bernard, O., Mitjavila, M. T., Romeo, P. H., Vainchenker, W., and Mathieu-Mahul, D., 1993. Expression of tal-1 and GATA-binding proteins during human hematopoiesis. *Blood*, **81**(3):647–655.

Murchison, E. P., Kheradpour, P., Sachidanandam, R., Smith, C., Hodges, E., Xuan, Z., Kellis, M., Grützner, F., Stark, A., and Hannon, G. J., *et al.*, 2008. Conservation of small RNA pathways in platypus. *Genome Research*, **18**(6):995 –1004.

Murchison, E. P., Tovar, C., Hsu, A., Bender, H. S., Kheradpour, P., Rebbeck, C. A., Obendorf, D., Conlan, C., Bahlo, M., Blizzard, C. A., *et al.*, 2010. The tasmanian devil transcriptome reveals schwann cell origins of a clonally transmissible cancer. *Science*, **327**(5961):84 –87.

Muto, A., Hoshino, H., Madisen, L., Yanai, N., Obinata, M., Karasuyama, H., Hayashi, N., Nakauchi, H., Yamamoto, M., Groudine, M., *et al.*, 1998. Identification of Bach2 as a B-cell-specific partner for small Maf proteins that negatively regulate the immunoglobulin heavy chain gene 3′ enhancer. *EMBO J*, **17**(19):5734–5743.

Nagarajan, P., Onami, T. M., Rajagopalan, S., Kania, S., Donnell, R., and Venkatachalam, S., 2009. Role of chromodomain helicase DNA-binding protein 2 in DNA damage response signaling and tumorigenesis. *Oncogene*, **28**(8):1053–1062.

Nascimento, E. M., Cox, C. L., Macarthur, S., Hussain, S., Trotter, M., Blanco, S., Suraj, M., Nichols, J., Kübler, B., Benitah, S. A., *et al.*, 2011. The opposing transcriptional functions of sin3a and c-Myc are required to maintain tissue homeostasis. *Nature Cell Biology*, **13**(12):1395–1405.

Nateri, A. S., Spencer-Dene, B., and Behrens, A., 2005. Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development. *Nature*, **437**(7056):281–285.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., Sandstrom, R., Johnson, A. K., Humbert, R., *et al.*, 2011. An expansive human regulatory lexicon encoded in transcription factor footprints. Submitted, NCP008.

Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M., 2003. Scanning human gene deserts for Long-Range enhancers. *Science*, **302**(5644):413.

Noyes, M. B., Christensen, R. G., Wakabayashi, A., Stormo, G. D., Brodsky, M. H., and Wolfe, S. A., 2008a. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**(7):1277–1289.

Noyes, M. B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M. H., and Wolfe, S. A., 2008b. A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Research*, **36**(8):2547–2560.

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E., *et al.*, 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, **39**(6):730–732.

O'Geen, H., Lin, Y., Xu, X., Echipare, L., Komashko, V. M., He, D., Frietze, S., Tanabe, O., Shi, L., Sartor, M. A., *et al.*, 2010. Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. *BMC Genomics*, **11**:689.

Park, P. J., 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**(10):669–680.

Partington, G. A. and Patient, R. K., 1999. Phosphorylation of GATA-1 increases its DNA-binding affinity and is correlated with induction of human K562 erythroleukaemia cells. *Nucleic Acids Research*, **27**(4):1168 –1175.

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., and Shendure, J., 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, **27**(12):1173–1175.

Paun, A. and Pitha, P. M., 2007. The IRF family, revisited. *Biochimie*, **89**(6-7):744–753.

Pavesi, G., Mauri, G., and Pesole, G., 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17**:S207–S214.

Peng, J., Zhu, Y., Milton, J. T., and Price, D. H., 1998. Identification of multiple cyclin subunits of human P-TEFb. *Genes & Development*, **12**(5):755–762.

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

Philippakis, A. A., Busser, B. W., Gisselbrecht, S. S., He, F. S., Estrada, B., Michelson, A. M., and Bulyk, M. L., 2006. Expression-Guided in silico evaluation of candidate cis regulatory codes for *Drosophila* muscle founder cells. *PLoS Computational Biology*, **2**(5):e53.

Phillips, J. E. and Corces, V. G., 2009. CTCF: master weaver of the genome. *Cell*, **137**(7):1194–1211.

Pietrokovski, S., 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, **24**(19):3836–3845.

Pop, M., Salzberg, S. L., and Shumway, M., 2002. Genome sequence assembly: algorithms and issues. *Computer*, **35**(7):47–54.

Pruitt, K. D. and Maglott, D. R., 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, **29**(1):137 –140.

Qi, Y., Rolfe, A., MacIsaac, K. D., Gerber, G. K., Pokholok, D., Zeitlinger, J., Danford, T., Dowell, R. D., Fraenkel, E., Jaakkola, T. S., *et al.*, 2006. High-resolution computational models of genome binding events. *Nature Biotechnology*, **24**(8):963–970.

Quinlan, A. R. and Hall, I. M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6):841 –842.

Quinlan, J. R., 1996. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, **4**:77–90. Journal of Artificial Intelligence Research, Vol 4, (1996), 77-90.

Raha, D., Wang, Z., Moqtaderi, Z., Wu, L., Zhong, G., Gerstein, M., Struhl, K., and Snyder, M., 2010. Close association of RNA polymerase II and many transcription factors with pol III genes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(8):3639–3644.

Reed, D. E., Huang, X. M., Wohlschlegel, J. A., Levine, M. S., and Senger, K., 2008. DEAF-1 regulates immunity gene expression in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(24):8351–8356.

Reis-Filho, J. S., Westbury, C., and Pierga, J., 2006. The impact of expression profiling on prognostic and predictive testing in breast cancer. *Journal of Clinical Pathology*, **59**(3):225 –231.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.*, 2000. Genome-Wide location and function of DNA binding proteins. *Science*, **290**(5500):2306 –2309.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al.*, 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, **4**(8):651–657.

Roder, K., Wolf, S., Larkin, K., and Schweizer, M., 1999. Interaction between the two ubiquitously expressed transcription factors NF-Y and sp1. *Gene*, **234**(1):61–69.

Roider, H. G., Manke, T., O'Keeffe, S., Vingron, M., and Haas, S. A., 2009. PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, **25**(4):435 –442.

Romer, K. A., Kayombya, G., and Fraenkel, E., 2007. WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and bayesian approaches. *Nucleic Acids Research*, **35**(Web Server):W217–W220.

Rothbächer, U., Bertrand, V., Lamy, C., and Lemaire, P., 2007. A combinatorial code of maternal GATA, ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development*, **134**(22):4023–4032.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor, G. L., Miklos, Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., *et al.*, 2000. Comparative genomics of the eukaryotes. *Science*, **287**(5461):2204 –2215.

Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., Aebersold, R., Ranish, J. A., and Krumm, A., 2008. CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(24):8309–8314.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P., 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, **127**(6):1193–1207.

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P., and Lai, E. C., 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Research*, **17**(12):000.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, **32**(Database issue):D91–94.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E. E., 2007. A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes & Development*, **21**(4):436–449.

Sandmann, T., Jensen, L. J., Jakobsen, J. S., Karzynski, M. M., Eichenlaub, M. P., Bork, P., and Furlong, E. E., 2006. A temporal map of transcription factor activity: Mef2 directly regulates target genes at all stages of muscle development. *Developmental Cell*, **10**(6):797–807.

Sanger, F., Nicklen, S., and Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**(12):5463 –5467.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235):467 –470.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., *et al.*, 2010. Five-Vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**(5981):1036–1040.

Schnall-Levin, M., Zhao, Y., Perrimon, N., and Berger, B., 2010. Conserved microRNA targeting in *Drosophila* is as widespread in coding regions as in 3′UTRs. *Proceedings of the National Academy of Sciences*, **107**(36):15751 –15756.

Schones, D. E., Sumazin, P., and Zhang, M. Q., 2005. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**(3):307 –313.

Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D., and Gaul, U., 2004. Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biology*, **2**(9):e271.

Schuster, S. C., 2008. Next-generation sequencing transforms today's biology. *Nature Methods*, **5**(1):16–18.

Scott, E., Simon, M., Anastasi, J., and Singh, H., 1994. Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science*, **265**(5178):1573 –1577.

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

Sementchenko, V. I. and Watson, D. K., 2000. Ets target genes: past, present and future. *Oncogene*, **19**(55):6533–6548.

Sen, A., Stultz, B. G., Lee, H., and Hursh, D. A., 2010. Odd paired transcriptional activation of decapentaplegic in the *Drosophila* eye/antennal disc is cell autonomous but indirect. *Developmental Biology*, **343**(1-2):167–177.

Seto, E., Shi, Y., and Shenk, T., 1991. YY1 is an initiator sequence-binding protein that directs and activates transcription in vitro. *Nature*, **354**(6350):241–245.

Sharan, R., Ovcharenko, I., Ben-Hur, A., and Karp, R. M., 2003. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, **19**(Suppl 1):i283–i291.

Shendure, J. and Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, **26**(10):1135–1145.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M., *et al.*, 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**(5741):1728 –1732.

Sobek-Klocke, I., Disqué-Kochem, C., Ronsiek, M., Klocke, R., Jockusch, H., Breuning, A., Ponstingl, H., Rojas, K., Overhauser, J., and Eichenlaub-Ritter, U., *et al.*, 1997. The human gene ZFP161 on 18p11.21-pter encodes a putative c-myc repressor and is homologous to murine zfp161 (Chr 17) and zfp161-rs1 (X chr). *Genomics*, **43**(2):156–164.

Solomon, M. J., Larsen, P. L., and Varshavsky, A., 1988. Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone h4 is retained on a highly transcribed gene. *Cell*, **53**(6):937–947.

Solozobova, V., Rolletschek, A., and Blattner, C., 2009. Nuclear accumulation and activation of p53 in embryonic stem cells after DNA damage. *BMC Cell Biology*, **10**(1):46.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., *et al.*, 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, **21**(10):1757–1767.

Spivakov, M., , Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Furlong, E., and Birney, E., *et al.*, 2011. Analysis of variation at transcription factor binding sites in *Drosophila* and humans. Submitted, GRCP027.

Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M., 2005. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, **123**(6):1133–1146.

Stark, A., Bushati, N., Jan, C. H., Kheradpour, P., Hodges, E., Brennecke, J., Bartel, D. P., Cohen, S. M., and Kellis, M., 2008. A single hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands. *Genes & Development*, **22**(1):8 –13.

Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G. J., and Kellis, M., 2007a. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Research*, **17**(12):1865–1879.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., *et al.*, 2007b. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, **450**(7167):219–232.

Sun, H., Yuan, Y., Wu, Y., Liu, H., Liu, J. S., and Xie, H., 2010. Tmod: toolbox of motif discovery. *Bioinformatics*, **26**(3):405 –407.

Suzuki, M. M. and Bird, A., 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, **9**(6):465–476.

Takahashi, K. and Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**(4):663–676.

Taylor, J. M., Dupont-Versteegden, E. E., Davies, J. D., Hassell, J. A., Houlé, J. D., Gurley, C. M., and Peterson, C. A., 1997. A role for the ETS domain transcription factor PEA3 in myogenic differentiation. *Molecular and Cellular Biology*, **17**(9):5550–5558.

Thompson, J. D., Higgins, D. G., and Gibson, T. J., 1994. CLUSTAL w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**(22):4673–4680.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S. E., *et al.*, 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biology*, **3**(12):research00881–8814.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., *et al.*, 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, **23**(1):137–144.

Touzet, H. and Varre, J., 2007. Efficient and accurate p-value computation for position weight matrices. *Algorithms for Molecular Biology*, **2**(1):15.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., *et al.*, 2009. FlyBase: enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Research*, **37**:D555–D559.

Tyler, D. M., Okamura, K., Chung, W., Hagen, J. W., Berezikov, E., Hannon, G. J., and Lai, E. C., 2008. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes & Development*, **22**(1):26 –36.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.*, 2001. The sequence of the human genome. *Science*, **291**(5507):1304 –1351.

Vicoso, B. and Charlesworth, B., 2006. Evolution on the x chromosome: unusual patterns and processes. *Nature Reviews Genetics*, **7**(8):645–653.

Villard, J., Peretti, M., Masternak, K., Barras, E., Caretti, G., Mantovani, R., and Reith, W., 2000. A functionally essential domain of RFX5 mediates activation of major histocompatibility complex class II promoters by promoting cooperative binding between RFX and NF-Y. *Molecular and Cellular Biology*, **20**(10):3364 –3376.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., *et al.*, 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**(7231):854–858.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A., 2007. VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Research*, **35**(Database issue):D88–92.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T., Greven, M., Aldred, S. F., Trinklein, N., Dong, X., Kundaje, A., *et al.*, 2011. Genome-wide mapping of the binding sites of 119 human transcription factors. Submitted, NCP010.

Wang, W., Xue, Y., Zhou, S., Kuo, A., Cairns, B. R., and Crabtree, G. R., 1996. Diversity and specialization of mammalian SWI/SNF complexes. *Genes & Development*, **10**(17):2117 –2130.

Wang, Z. and Burge, C. B., 2008. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**(5):802 –813.

Wang, Z., Gerstein, M., and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1):57–63.

Ward, L. D. and Bussemaker, H. J., 2008. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics*, **24**(13):i165 –i171.

Wasserman, W. W. and Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, **5**(4):276–287.

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., Losick, R., and CSHLP, I., 2007. *Molecular Biology of the Gene*. Benjamin Cummings, 6 edition.

Watson, J. D. and Crick, F. H. C., 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, **171**(4356):737–738.

Wei, C., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., *et al.*, 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**(1):207–219.

Weiss, M. and Orkin, S., 1995. GATA transcription factors: key regulators of hematopoiesis. *Experimental hematology*, **23**(2):99.

Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., *et al.*, 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, **451**(7180):796–801.

Wharton, R. P., Sonoda, J., Lee, T., Patterson, M., and Murata, Y., 1998. The pumilio RNA-Binding domain is also a translational regulator. *Molecular Cell*, **1**(6):863–872.

Wilson, E. B., 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**(158):209–212.

Winter, J., Jung, S., Keller, S., Gregory, R. I., and Diederichs, S., 2009. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature Cell Biology*, **11**(3):228–234.

Xi, H., Shulha, H. P., Lin, J. M., Vales, T. R., Fu, Y., Bodine, D. M., McKay, R. D. G., Chenoweth, J. G., Tesar, P. J., Furey, T. S., *et al.*, 2007. Identification and characterization of cell Type-Specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genetics*, **3**(8):e136.

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

Xie, X., Rigor, P., and Baldi, P., 2009. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, **25**(2):167–174.

Xu, D., Zhao, L., Del Valle, L., Miklossy, J., and Zhang, L., 2008. Interferon regulatory factor 4 is involved in Epstein-Barr virus-mediated transformation of human b lymphocytes. *Journal of Virology*, **82**(13):6251–6258.

Yi, F. and Merrill, B. J., 2007. Stem cells and TCF proteins: A role for Beta-Catenin-Independent functions. *Stem Cell Reviews*, **3**(1):39–48.

Yu, H., Mashtalir, N., Daou, S., Hammond-Martel, I., Ross, J., Sui, G., Hart, G. W., Rauscher, Frank J, r., Drobetsky, E., Milot, E., *et al.*, 2010. The ubiquitin carboxyl hydrolase BAP1 forms a ternary complex with YY1 and HCF-1 and is a critical regulator of gene expression. *Molecular and Cellular Biology*, **30**(21):5071–5085.

Yu, L., Wu, Q., Yang, C. P., and Horwitz, S. B., 1995. Coordination of transcription factors, NF-Y and C/EBP beta, in the regulation of the mdr1b promoter. *Cell Growth & Differentiation*, **6**(12):1505–1512.

Zaret, K. S. and Carroll, J. S., 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, **25**(21):2227 –2241.

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A., and Levine, M., 2007. Whole-genome ChIP-chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes & Development*, **21**(4):385–390.

Zerbino, D. R. and Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, **18**(5):821 –829.

Zervos, A. S., Gyuris, J., and Brent, R., 1993. Mxi1, a protein that specifically interacts with max to bind Myc-Max recognition sites. *Cell*, **72**(2):223–232.