DNA SEQUENCE AND EPIGENETIC FEATURES THAT DRIVE
HUMAN PROMOTER FUNCTION

by

JANE MARIE LANDOLIN

B.S., University of California, Los Angeles, 2003
M.S., University of California, Los Angeles, 2003

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

2009
Approved by

First Reader

Zhiping Weng, Ph.D.
Professor of Biochemistry and Molecular Pharmacology,
University of Massachusetts Medical School;
Adjunct Professor of Biomedical Engineering, Boston University

Second Reader

Mark A. Kon, Ph.D.
Professor of Mathematics and Statistics

Third Reader

Yu (Brandon) Xia, Ph.D.
Assistant Professor of Chemistry

Fourth Reader

Tom D. Tullius, Ph.D.
Professor of Chemistry

Fifth Reader

Charles DeLisi, Ph.D.
Metcalf Professor of Science and Engineering
Acknowledgments

This dissertation is the culmination of five long years in Boston, Massachusetts; three thousand miles away from the last home I knew, and half way around the globe from where I grew up. I would not have embarked on this journey if it were not for my father Jack Lin, who decided long ago that his children would be educated in the United States. Eventually, I made my way to the University of California in Los Angeles, where my undergraduate advisor Joseph DiStefano III took me under his wing. He encouraged all of his students to pursue what they loved and love what they pursued. He was a mentor in both science and in life, and I am profoundly grateful to have known him.

I am deeply indebted to my graduate advisor Zhiping Weng, who provided intellectual support and resources for the work presented in this dissertation. Most notably, she offered me a rare opportunity to participate in “big science” through the ENCODE project. Research in Zhiping’s lab was trial by fire; she was intolerant of sloppy thinking and incoherent results, and I quickly learned to sharpen my skills under her direction. All of her graduate students would come to realize her generosity, and the incredible value of both tangible and intangible lessons she teaches. I am grateful to have trained with one of the most accomplished genomic scientists of our time.

I would like to thank Kenneth Lutchen, former chair of Biomedical Engineering, who secured a grant from the National Institutes of Health for graduate training in Quantitative Biology and Physiology that supported my first two years at Boston University. I thank my dissertation committee members, Yu Xia, Tom Tullius, Charles DeLisi. In particular, Mark Kon provided guidance on the computational details of implementing support vector machines, and Tom Tullius helped with editorial changes that were much needed to refine the final draft of this dissertation.

I would like to acknowledge my colleagues and fellow data-wranglers in Zhiping’s lab, especially Anason Halees, Yutao Fu, Ulas Karaöz, and Soo Lee, who helped me compile my bag of bioinformatics tips and tricks. I thank Mary-Ellen Fitzpatrick, our proficient Net-
work/Systems Administrator, who kept our computers functioning despite unintentional (and intentional) assaults to the operating system. I was fortunate to collaborate with superb experimentalists from Richard Myers’ laboratory at the Department of Human Genetics, at Stanford University. Nathan Trinkelin and Patrick Collins, now with the company Switchgear Genomics, performed promoter assays and Chromatin immuno-precipitation (ChIP) experiments for the bidirectional promoters project, and mutagenesis for the transcription factor footprint project. David Johnson, now with the company Gene Security Network, oversaw high-throughput promoter assays and gene expression experiments for the study on tissue-specific regulation. Alayne Browne performed ChIP-seq experiments studying DNA methylation. None of this work would have been possible without their quick minds and fierce pipetting skills.

I thank my friends Melissa Landon and Joseph Mellor, first for stimulating discussions; I always generate the best ideas after chatting with them, and second, for a place to stay; this dissertation was largely written during an ascetic tenure in their basement. I would not have survived graduate school without Jaafar Haidar, Henry Lee, Hemali Patel, Sunmin Ahn, Lauren Hayward, Jessica Gereige, Patrick Allen, and Barrett Nehilla who made my work better by occasionally accompanying me in play.

My deepest gratitude goes to my family for their constant support and confidence in my abilities. I thank my sister Jo-Ann Lin for helping me be a better sister, daughter, and scientist by reminding me of the important things in life. I thank my mother Mary Lee, for her intelligence and good humor, and my father Jack Lin, for his diligence and principled thinking.

I especially thank my wife Chelsea Landolin whose patient love enabled me to complete this work.
DNA SEQUENCE AND EPIGENETIC FEATURES THAT DRIVE
HUMAN PROMOTER FUNCTION

(Order No. )

JANE MARIE LANDOLIN

Boston University, College of Engineering, 2009

Major Professor: Zhiping Weng, Ph.D.,
Professor of Biochemistry and Molecular Pharmacology,
University of Massachusetts Medical School;
Adjunct Professor of Biomedical Engineering,
Boston University

ABSTRACT

The human genome contains approximately three billion bases that instruct cells to produce proteins. Varying levels of proteins inside of cells affect the function of tissues and organs, and eventually influence the health and behavior of individuals. One of the first steps in this chain of events happens at the molecular level, when DNA is transcribed into RNA. Promoters contain the necessary and sufficient information for cells to initiate transcription. Interspersed along promoters are 6 to 10 base sequence motifs that are bound by transcription factors (TFs), which activate or repress transcription of downstream genes. Epigenetic marks such as CpG methylation and modified histones can also affect transcription by interacting with TFs. As the sequence of the Human genome is now known, and experimental techniques to map epigenetic marks genome-wide are now being attempted, we have an unprecedented opportunity to study of transcription regulation with both breadth and depth.

In this dissertation, I present several studies examining regulatory signals from DNA sequence as well as epigenetic marks. The first study identifies TFs and modified histones that are important for regulating bidirectional promoters in the human genome. The second study models tissue-specific transcription regulation using TF sequence motifs and CpG
methylation. Building upon the biological principles discovered in the first two studies, the third study computationally predicts TF binding footprints, which are then tested experimentally by mutagenesis.

GA-binding protein (GABP) is an important TF identified in all three studies. The analysis presented in this dissertation shows that GABP is a major regulator of bidirectional promoters, and activates genes that are ubiquitously transcribed. The motif for GABP is a good predictor of in vivo binding, and mutagenesis confirms that our TF footprint predictions are indeed functional, with average 3-fold knock down of promoter activity compared to wild-type.

The analysis was possible because DNA sequence signals were decoupled from epigenetic marks using a transient transfection promoter activity assay system. This work demonstrates the importance of computation in analyzing biological data from different types of high-throughput experiments, and provides the foundation for a single-base-resolution map of TF binding footprints in the Human genome.
## Contents

1 Background  
1.1 DNA sequence and epigenetic regulators of transcription . . . . . . . . . . . . 1  
1.2 Computational models of TF binding . . . . . . . . . . . . . . . . . . . . . 2  
1.3 Experimental strategies to map TF binding sites . . . . . . . . . . . . . . 6  
1.4 The transient transfection system . . . . . . . . . . . . . . . . . . . . . . . 7  

2 Bidirectional Promoters in the Human Genome  
2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10  
2.2 \textit{Ab initio} motif discovery in bidirectional promoters . . . . . . . . 12  
2.3 TRANSFAC motif presence in bidirectional promoters . . . . . . . . . . . 13  
2.4 Positional preference of motifs in bidirectional promoters . . . . . . . . . 15  
2.5 Divergent genes are co-regulated . . . . . . . . . . . . . . . . . . . . . . . 17  
2.6 TF binding in bidirectional promoters . . . . . . . . . . . . . . . . . . . . 17  
2.7 Modified histones around bidirectional promoters . . . . . . . . . . . . . . 20  
2.8 A case study of GABPA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22  
2.8.1 GABPA binds to a majority of bidirectional promoters . . . . . . . . 22  
2.8.2 GABPA motif presence correlates with binding in living cells . . . . 24  
2.8.3 GABPA-binding sites often required for bidirectional activity . . . . 26  
2.9 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29  
2.9.1 Two bidirectional promoter data sets . . . . . . . . . . . . . . . . . . . 29  
2.9.2 One unidirectional promoter data set . . . . . . . . . . . . . . . . . . . 30  
2.9.3 Gene ontology analysis . . . . . . . . . . . . . . . . . . . . . . . . . . 30  
2.9.4 Motif discovery . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30  
2.9.5 Motif representation calculation . . . . . . . . . . . . . . . . . . . . . 31
2.9.6 ENCODE ChIP-chip factors in bidirectional promoters 32
2.9.7 Histone modification analysis 32
2.9.8 Microarray analysis 33
2.9.9 Deletion constructs 33
2.9.10 Mutation analysis of 30-bp fragment 33
2.9.11 GABPA ChIP-qPCR 34

3 Tissue-specific Transcription Regulation 36
3.1 Introduction 36
3.2 Promoter activity and gene expression data 39
3.3 Genetic and epigenetic contributions to endogenous expression 41
3.4 Novel promoters 43
3.5 Predictive models of promoter activity 46
  3.5.1 Exploratory analysis of motif representation 46
  3.5.2 Exploratory analysis of modeling strategies 47
3.6 The effect of CG dinucleotide 50
3.7 Predicting ubiquitous and cell line-specific active promoters 51
3.8 Cis-regulatory module discovery 55
3.9 Predicting ubiquitous and cell line-specific expressed genes 55
3.10 Methods 57
  3.10.1 Promoter transient transfection reporter assay 57
  3.10.2 Whole genome expression analysis 59
  3.10.3 CG dinucleotide count normalization 59
  3.10.4 SVM implementation details 59
  3.10.5 Reducing redundancy in motif sets 60

4 Single-base-resolution Map of TF Footprints in Human Promoters 61
4.1 Introduction 61
4.2 TF binding footprints in Human promoters 62
4.3 Filtering criterion 1: TF binding site must be solitary .......................... 63
4.4 Filtering criterion 2: The highest scoring TF binding site must be statistically significant .......................................................... 63
4.5 Preliminary mutagenesis of five TF footprints ................................. 65
4.6 Possible filtering criterion: Conservation ........................................ 65
4.7 Possible filtering criterion: Promoter overlap with ChIP target lists .... 69

5 Conclusion .................................................................................. 71
  5.1 Summary .............................................................................. 71
  5.2 Future directions ................................................................. 72
    5.2.1 Integrated models of transcription regulation ....................... 72
    5.2.2 Synthetic promoters .................................................... 72

References .................................................................................... 74

Curriculum Vitae ............................................................................. 91
List of Tables

3.1 AUCs of best performing models ........................................ 53
3.2 Motif modules discovered using the promoter activity dataset ............. 56
3.3 Motif modules discovered using the gene expression dataset ................. 58
### List of Figures

1-1  Representations of transcription factor binding sites .......................... 5
1-2  Chromatin immunoprecipitation (ChIP) methods ................................. 8

2-1  Three types of motif representation in bidirectional promoters ............. 11
2-2  Sequence logos of unique motifs ......................................................... 14
2-3  Enrichment of motifs, transcription factors, and modified histones ........ 16
2-4  Position specificity histograms of \emph{ab initio} motifs .......................... 18
2-5  The distribution of Pearson correlation coefficients ............................. 19
2-6  Fraction of promoters overlapping ChIP-chip target lists .................... 21
2-7  Histone modification ChIP intensities downstream of ENCODE genes ...... 22
2-8  Discretized histone modification patterns are not uniformly distributed. .. 23
2-9  GABPA ChIP in three region categories and four motif score cutoffs ....... 25
2-10 GABPA binding site presence ............................................................. 27
2-11 A deletion construct experiment with mapped TF binding sites ............ 28
2-12 Mutagenesis of a 30-bp bidirectional fragment ................................. 29

3-1  Distribution of promoter activity and gene expression scores ................. 41
3-2  Binarized promoter activity patterns .................................................... 42
3-3  Fraction of ubiquitous promoters and genes .......................................... 44
3-4  Fraction of novel and known active promoters ....................................... 45
3-5  Linear relationship between GABPA sites and ChIP fold enrichment ....... 48
3-6  Prediction accuracies among six machine-learning algorithms ............... 49
3-7  Distribution of CG dinucleotide content in 4575 promoters. ................ 52
3-8  Comparison of TF feature selection by RSVM and univariate ranking. .... 54
4.1 Distribution of GABPA Possum scores in Human promoters . . . . . . . . . . 64
4.2 Mutagenesis of five GABPA binding footprints . . . . . . . . . . . . . . . . 66
4.3 Conservation of a GABPA binding footprint (CHR2_P1199_R1). . . . . . 67
4.4 Conservation of a GABPA binding footprint (CHR7_M0172_R1). . . . . . 67
4.5 Conservation of a GABPA binding footprint (CHR21_M0109_R1). . . . . . 68
4.6 Conservation of a GABPA binding footprint (CHR22_P0237_R3). . . . . . 68
4.7 Conservation of a GABPA binding footprint (CHR_X_P0696_R1). . . . . . 69
4.8 Overlap of Human promoters with ChIP target lists. . . . . . . . . . . . . . 70
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Area Under the ROC Curve</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin Immuno-Precipitation</td>
</tr>
<tr>
<td>CRM</td>
<td>Cis-Regulatory Module</td>
</tr>
<tr>
<td>ENCODE</td>
<td>ENCyclopedia of DNA Elements</td>
</tr>
<tr>
<td>GABPA</td>
<td>GA-Binding Protein α</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NHGRI</td>
<td>National Human Genome Research Institute</td>
</tr>
<tr>
<td>PET</td>
<td>Paired End Tags</td>
</tr>
<tr>
<td>PSSM</td>
<td>Position Specific Scoring Matrix</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>S/N</td>
<td>Signal-to-Noise ratio</td>
</tr>
<tr>
<td>STAGE</td>
<td>Sequence Tag Analysis of Genomic Enrichment</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription Start Site</td>
</tr>
</tbody>
</table>
Chapter 1

Background

1.1 DNA sequence and epigenetic regulators of transcription

Gene regulation largely determines the morphology, function, and dynamic response of cells (Kadonaga, 2004; Levine and Tjian, 2003; Mitchell and Tjian, 1989). Nuclear proteins recognize and bind to DNA sequence elements, initiating a chain of events that allows a variety of cell types to perform specialized functions in order to maintain homeostasis. The human body, for example, contains approximately 210 different cell types with dramatically different phenotypes (Su et al., 2002).

Transcription begins as RNA polymerase II recognizes and binds to the transcription start sites (TSS) of genes. Transcription factors (TFs) further fine-tune the process according to environmental stimuli or dynamics of the cell cycle (Gacto et al., 2003; Kadonaga, 2004). TFs drive tissue specificity by activating transcription of genes that are unique to certain tissues. HNF4α, for instance, binds to promoter regions of liver-specific genes such as OCT1, which mediates hepatocellular uptake of organic cations critical for eliminating amines, drugs, and toxins (Saborowski et al., 2006). Many TFs are sequence-specific, recognizing and binding to particular sequences of DNA. HNF4α recognizes the consensus CAAAGG, while GABPA recognizes the consensus CCGGAA.

By contrast, epigenetic regulators of transcription are not informed by sequence, and can have complex interactions with TFs. DNA methylases add a methyl (−CH$_3$) group to cytosines that form CG dinucleotides, generally repressing transcription (Vaniushin, 2006). In promoters, TF binding sites can be methylated, changing the biochemical signature around binding sites such that TFs can no longer bind, resulting in repressed transcription.
(Vaniushin, 2006; Weber et al., 2007). Chromatin state can also block TF access to binding sites when DNA is tightly wrapped around histones. In response, some transcription factors couple with chromatin remodeling domains to move and modify histone proteins (Ito et al., 1997; Croston and Kadonaga, 1993). Histone tails can be modified by methylation, acetylation, and ubiquination, with varying consequences on transcription (Pazin and Kadonaga, 1997; Turner, 2002). Tri-methylation of the fourth lysine residue in the third histone (H3K4me3) is associated with silencing transcription, while acetylated histones in the gene body is a mark for actively transcribed genes (Turner, 2005). Every transcription event is therefore orchestrated by a complex interplay between DNA sequence and epigenetic regulators.

1.2 Computational models of TF binding

For the most part, computational models to predict TF binding sites have ignored epigenetic interactions, instead focusing only on DNA sequence input. As the genomes of many species have now been sequenced (Lander et al., 2001; Waterston et al., 2002; Kirkness et al., 2003; Gibbs et al., 2004; Sequencing and Consortium, 2005), one can theoretically predict all TF binding sites by searching for occurrences of consensus motifs in sequenced genomes. The reality, however, is that the majority of these computational predictions are false positives - they do not bind in vivo, or are non-functional - they bind, but have no transcriptional consequence. To illustrate this point, consider the expectation of observing any given motif of length eight. By chance alone, one would observe $3 \times 10^9 / 4^8 = 45,776$ instances of this motif in the human genome, containing about 3 billion base pairs. Even if a TF binds once to all 20,000 genes in the human genome, it will still have to contend with a background noise that is more than twice as frequent. At best, c-Myc binds to 46,000 locations in ST2091 cells ($S/N = 1$), at worst Stat1 binds to only 2,500 locations in HeLa cells ($S/N = 0.0546$) (Birney et al., 2007). Therefore, the presence of a consensus motif is only an indicator for potential binding, and in vivo binding is, in addition, governed by the epigenetic state of DNA around the motif.
To improve \textit{in silico} predictions and discover biological signals in a sea of noise, computational biologists have further developed algorithms to take into account cross-species conservation and cooperative interactions among TFs. As regulatory sequences tend to be conserved across species, limiting predictions to conserved blocks can eliminate spurious results. The programs rVista and ConSite filtered 90\% of their results, according to multi-species alignments (AVID and OCRA algorithms respectively), while maintaining 70\% sensitivity (Wasserman and Sandelin, 2004). The group of Manolis Kellis at MIT produced a compendium of TF binding sites in human promoters using a scoring scheme that takes into account the evolutionary distances among vertebrate species (Xie et al., 2005).

Another way to improve \textit{in silico} predictions is to look for clusters of motifs, called cis-regulatory modules (CRMs). The assumption is that TFs will interact and work together to coordinate transcription. Groups of co-regulated genes, as determined by Microarray Gene Expression experiments or Gene Ontology, are first stratified. CRM algorithms scan the promoter sequences for TF combinations with a common regulatory circuit or logic. These algorithms come in various flavors. Cluster-buster (Frith et al., 2003), for example, looks for motif clusters according to a gap parameter that specifies the optimal distance among motifs within a cluster. Cayce (Lin and Weng, 2006) finds statistically overrepresented motif association rules. MotifRegressor (Conlon et al., 2003) and LogicMotif (Keles et al., 2004) find interaction rules by linear and logic regression, while Beer and Tavazoie predict gene expression profiles in \textit{S. cerevisiae} and \textit{C. elegans} by inferring regulatory rules using Bayesian networks (Beer and Tavazoie, 2004). Complex network models such as REDUCE (Roven and Bussemaker, 2003) and CLR (Faith et al., 2007) incorporate the expression levels of TFs, but do not map binding sites per se. Newer algorithms combine both phylogenetic information and CRM search ( Marinescu et al., 2005), and meta-algorithms can additionally incorporate genome wide ChIP-chip data. This combination strategy appears to be effective at discriminating bound sites from false predictions with an accuracy of up to 80\% (Harbison et al., 2004; Holloway et al., 2005). As high-throughput epigenetic exper-
iments are now feasible to attempt, we will likely see expansions of the existing algorithms to incorporate new types of data.

Motif finding algorithms quantify TF binding sites in two main ways. The first way is to define a consensus sequence composed of the four nucleotides, which can be extended to accept degeneracy: Y for pyrimidine (C or T); R for purine (A or G); W for weak (A or T); S for strong (C or G); and N for any (A, C, G, or T). This model is used in the motif finding algorithms YMF (Sinha and Tompa, 2003), FindExplanators (Blanchette and Sinha, 2001), DME (Smith et al., 2005), Consensus (Ulyanov and Stormo, 1995) and also used to represent conserved motifs by Xie and Kellis (Xie et al., 2005). A more sophisticated way to quantify TFBS is to use 4-by-N position specific scoring matrices (PSSM), where four rows represent the four possible nucleotides, and N columns span the TF footprint. A raw count matrix can be obtained by aligning TF footprints from several binding experiments, and counting the number of each nucleotide at each position (Fig. 1.1c). This matrix can be further converted to frequencies or probabilities by dividing by the total number of binding experiments, or converted to log-likelihood weights by taking the log of the frequencies and comparing them to appropriate background nucleotide frequencies (Fig. 1.1d). In the motif scanning algorithm POSSUM, a fragment of DNA will be assigned a high score if it closely resembles the log-likelihood model defined by a given PSSM, and will be given a low score if it resembles a model based on the nucleotide frequencies of a predefined set of background sequences. This scoring scheme is based on statistical-mechanical theory, and has been shown to correlate with measured binding affinities (Berg and von Hippel, 1987; Stormo, 2000). Motif finding algorithms such as MEME (Bailey and Gribskov, 1998), Gibbs Motif Sampler (Thompson et al., 2003), and Glam (Frith et al., 2004b) all provide PSSMs in their final output. Additionally, several public databases that compile TFBS annotations also use PSSMs as quantitative models: TRANSFAC (Heinemeyer et al., 1999; Matys et al., 2006), TRRD (Kolchanov et al., 2002), COMPEL (Matys et al., 2006), JASPAR (Vlieghe et al., 2006), IMD (Chen et al., 1995), and EPD (Schmid et al., 2006).
Figure 1-1: Transcription factor binding sites can be represented in many ways. a) Alignment of 8 MEF-2 binding site sequences. b) MEF-2 consensus sequence. c) MEF-2 position specific frequency or count matrix. d) MEF-2 position specific weight matrix, a.k.a. position specific scoring matrix. e) Binding site scores for each position of a representative binding sequence. f) Sequence logo representation. Image from (Wasserman and Sandelin, 2004)
1.3 Experimental strategies to map TF binding sites

Experimental mapping of TF binding sites has traditionally been performed in a low-throughput manner. DNA mobility shift assays can detect protein-bound DNA because the larger molecular weight impedes mobility through a gel when compared to unbound DNA (Fried and Crothers, 1981; Garner and Revzin, 1981). In DNase footprinting assays, the exact location of TF binding can be identified by a large gap in the DNA ladder on a polyacrylamide gel, corresponding to the uncleaved portion of the protein-bound DNA (Dynan and Tjian, 1983; Galas and Schmitz, 1978). The drawbacks of these methods are that they require DNA to be out of its native environment, and are not amenable for high-throughput testing.

Chromatin immunoprecipitation does not suffer from these drawbacks because the chromatin marks on DNA are undisturbed, and the DNA is in its native environment when it is mixed with the TF-specific antibody (Kirmizis and Farnham, 2004; Weinmann and Farnham, 2002; Weinmann, 2004). Formaldehyde crosslinking affixes the TF to DNA, and the DNA is randomly sheared. Antibodies that recognize the TF are added, which guide precipitation of the TF-bound DNA. The crosslinks are now removed, and the DNA can be treated in three ways to identify TF binding regions: 1) polymerase chain reaction (PCR), 2) hybridization to Microarrays, or 3) cloning and sequencing (Fig. 1.2). PCR methods can accurately amplify a TF-bound DNA fragment for detection, but require laborious design of PCR primers. This limits their usefulness to validation of hypothesized binding sites rather than discovery of novel sites. In Microarrays, the scope of detectable DNA fragments depends mainly on the design of the chip. An unbiased design is the tiling array, where non-repetitive genomic regions are laid out in order across the rows of a Microarray (Kapranov et al., 2002). This method has been applied for Sp1, cMyc and p53 in human chromosomes 21 and 22 (Cawley et al., 2004). Other Microarray designs, such as custom PCR product arrays, have also been successfully applied to find TF binding sites (Ren et al., 2000). Sequencing methods sidestep unnecessary sequencing of the entire precipi-
tated fragment by only sequencing tags demarcating the beginning and end of a fragment. These tags are then computationally mapped to regions in a reference genome for localization of a binding region. In sequence tag analysis of genomic enrichment (ChIP-STAGE), fragments are amplified by biotinylated primers, cut by restriction enzymes Mmel and NlaIII, linked, and concatenated to other tags before sequencing (Kim and Iyer, 2005).

In paired end ditag sequencing (ChIP-PET), fragments are first cloned into plasmid vectors before cutting by restriction enzyme and concatenation for sequencing (Wei et al., 2006). Sequencing methods have the added benefit of being able to detect TF binding sites in repetitive regions of the genome. Indeed, 25% of p53 binding sites identified by a recent ChIP-PET study mapped to multiple locations on the human genome, possibly corresponding to repetitive regions (Wei et al., 2006). With the availability of many reference genomes, and the reduced cost and increased speed of modern sequencers, the cost for ChIP-sequencing methods can be significantly reduced. Using next-generation sequencing technology, scientists have mapped unbiased genomic locations for NRSF (Johnson et al., 2007), STAT1 (Robertson et al., 2007), and a monumental compendium of 21 histone marks (Barski et al., 2007).

1.4 The transient transfection system

The experimental methods described above can identify regions of TF binding, but do not annotate their functional consequence, such as activation or repression of a downstream gene. Additionally, ChIP is limited by the availability of an antibody, and by a relatively coarse resolution (>1kb), considering TF binding sites are typically 6-10 base pairs long.

The transient transfection assay is a molecular cloning technique where a promoter fragment is cloned upstream of a reporter gene, for example luciferase, and the activity of the gene can be quantified by the level of luminescence. Molecular cloning was first developed in the 1970s and has become common practice in molecular biology labs today. The lab of Richard Myers at Stanford University has dramatically increased the throughput of molecular cloning, and turned what used to take weeks for a single clone, into a high-
Figure 1.2: Schematic of chromatin immunoprecipitation (ChIP) methods a) Precipitated fragments can be identified in three ways: 1) polymerase chain reaction (PCR), 2) cloning and sequencing, and 3) hybridization to Microarrays. Image from (Weinmann, 2004)
throughput molecular assay system where hundreds of clones can be made in a reasonable time (Trinklein et al., 2003). Cooper et al. systematically cloned 642 promoters in 1% of the human genome, and tested their activities in 16 cell lines (Cooper et al., 2006). Forty-five promoters were randomly selected for deletion construct analysis, where the upstream region of the promoters were truncated in a series of nested deletions (at -1 kb, -500 bp, -350 bp, -200 bp, -90 bp, and -40 bp). Each construct was then tested for promoter activity. Of the 40 bp fragments, 68% showed significant promoter activity over the background, indicating that the core promoter, which is just large enough to support the entire footprint of the transcription machinery, is frequently sufficient to initiate transcription. The benefit of the transfection assay system is that the cumulative effects of functional elements within each deletion construct can be measured. They observed that, on average, promoter activity increased as the length of the constructs increased, peaking at 350 bp, and then decreasing for constructs that were 500 and 1 kb long. On average, there was a reduction in activity of the larger fragments, although a range of behaviors was observed for individual promoters. In the SPAG4 promoter, for example, deletion analysis confirmed the presence of negative elements in the 372-295 bp and 898-372 bp upstream region of the SPAG4 promoter. This shows that the transient transfection assay system can discern the function of a small fragment of DNA, and deletion analysis can systematically hone in on even smaller fragments.
Chapter 2

Bidirectional Promoters in the Human Genome

2.1 Introduction

Mammalian genomes are highly complex, with neighboring genes arranged in divergent, convergent, tandem, anti-sense, and interleaving fashions (Carninci et al., 2005; Carninci, 2006; Carninci et al., 2006; Engstrom et al., 2006; Kapranov et al., 2005). Despite the vast genomic space, a substantial portion of human genes (11%) are arranged in a divergent, head-to-head fashion and controlled by relatively short bidirectional promoters (Trinklein et al., 2004). In this study, a bidirectional promoter is defined as an intergenic region that is <1 kb long, and flanked by the transcription start site (TSS) of a plus-strand gene on one side and the TSS of a minus-strand gene on the other. This abundance has been observed across several mammalian genomes (Li et al., 2006), suggesting that there is evolutionary pressure for conserving this type of gene-pair structure. Some divergent genes are related by function, in particular DNA repair (Adachi and Lieber, 2002; Trinklein et al., 2004), and could be coregulated in a way that takes advantage of their paired arrangement. Indeed, the expression patterns of divergent gene pairs are more correlated than those of randomly paired genes (Trinklein et al., 2004).

Investigators have examined the sequences of bidirectional promoters for clues as to how they regulate both downstream genes. For example, most bidirectional promoters lack TATA boxes and are both GC-rich and enriched in CpG islands (Adachi and Lieber, 2002; Takai and Jones, 2004; Trinklein et al., 2004). They also display a mirror sequence composition, such that Gs and Ts dominate on one side of the midpoint, while Cs and As dominate on the other side (Engstrom et al., 2006). How these general features in-
1. Overrepresented

2. Shared

3. Underrepresented

Example forward unidirectional promoter

Example reverse unidirectional promoter

Figure 2.1: Three types of motif representation in bidirectional promoters. Motifs in the overrepresented category occur more often (e.g., four times) than the sum of occurrences in two unidirectional promoters (e.g., $2 + 1 = 3$ times). Motifs in the shared category occur as often (e.g., three times) as the sum of occurrences in two unidirectional promoters. Motifs in the underrepresented category occur less often (e.g., once) than the sum of occurrences in two unidirectional promoters.

fluence the binding of transcription factors (TFs) to bidirectional promoters is unknown. While numerous studies have summarized the composition of sequence motifs in unidirectional promoters (Ettwiller et al., 2005; Xie et al., 2005), none have addressed bidirectional promoters specifically. TF-binding sites of individual bidirectional promoters have been experimentally identified (Bush et al., 2003; Carter and Avadhani, 1994; Patton et al., 2006), but the results are difficult to generalize. In this study, I formulate a three-category conceptual framework to classify sequence motifs in bidirectional promoters computationally as overrepresented, shared, or underrepresented (Fig. 2.1).

Because bidirectional promoters have the ability to activate two downstream genes concurrently, one of my goals was to identify sequence signatures that distinguish this mode
of regulation from that of unidirectional promoters. I found TFs and modified histones that were overrepresented in bidirectional promoters compared to unidirectional promoters. One of these was the transcription factor GABPA, and in collaboration with Richard Myers’ group at Stanford University, we designed several experiments to look closer at GABPA function. Patrick Collins experimentally verified the binding of GABPA in bidirectional promoters, and Nathan Trinkelin validated the function of a GABPA binding site in a 30 bp bidirectional promoter fragment. Detailed analysis reveal that the dominant mode of transcription in bidirectional promoters is to activate both downstream genes, and that this can be achieved with as little as a single GABPA binding site.

2.2 Ab initio motif discovery in bidirectional promoters

I first determined whether bidirectional promoters distinguish themselves from unidirectional promoters by housing a special set of motifs. I compiled two data sets of bidirectional promoters and one data set of unidirectional promoters for comparison (see Methods for details). The smaller, higher quality set of 376 bidirectional promoters was analyzed by the ab initio motif discovery algorithm MEME (Grundy et al., 1996), and the sequences from a larger set of 1304 bidirectional promoters were used to categorize motifs as overrepresented, shared, or underrepresented. I randomly sampled 1304 pairs of unidirectional promoters from a total of 13,205 in the human genome and matched them with the 1304 bidirectional promoters for CpG content and length to avoid selection bias. I required MEME to output 15 motifs, but several of these were reverse complements of each other or single nucleotide repeats, and thus, the list was reduced to seven unique motifs (Fig. 2.2). Among these, I categorized five as overrepresented (NRF-1, CCAAT, GABPA, YY1, and ACTACAnnTCCC), one as shared (SP1), and one as underrepresented (Novel 1) (Fig. 2.3). The increased presence of these motifs in bidirectional promoters is consistent with the overrepresentation of certain Gene Ontology (GO) categories among divergent genes (Boyle et al., 2004; Harris et al., 2004). For example, NRF-1 is a key regulator of nuclear genes encoding components of the mitochondrial transcription and replication machinery.
(Gopalakrishnan and Scarpulla, 1995), and divergent genes are enriched in the GO term "mitochondrion” (GO:0005739, P = 1.2 x 10^{-9}). Overexpression of YY1 has been shown to accelerate DNA repair (Oei and Shi, 2001) and there is also an enrichment for the molecular function "response to DNA damage stimulus” (GO:0006974, P = 8.2 x 10^{-6}) among divergent genes. The motif for GABPA (also called NRF-2) is also categorized as overrepresented, and like NRF-1, is implicated in the transcriptional regulation of several subunits of mitochondrial enzymes (Scarpulla, 2006). I will revisit GABPA in a case study further below.

CCAAT-boxes (bound by NF-Y) are frequently found in TATA-less promoters (Mantovani, 1999), and TATA boxes are underrepresented in bidirectional promoters (Trinklein et al., 2004). I also discovered a motif with the consensus ACTACAnnTCCC and classified it as overrepresented in bidirectional promoters. This motif was previously reported by Xie et al. with the consensus ACTAYRnnnCCCR and was ranked fourth of 174 motifs in terms of conservation across several mammalian promoters (Xie et al., 2005). Because the top three motifs in their study correspond to known transcription factors, ACTACAnnTCCC is actually the highest-ranking novel motif in their set. Given the enrichment of this motif in bidirectional promoters and its strong evolutionary conservation across mammalian promoters, its cognate transcription factor likely plays an important role in regulating bidirectional promoters.

All five overrepresented motifs in bidirectional promoters are among the most conserved motifs in mammalian promoters at large: NRF-1, CCAAT, GABPA, YY1, and ACTACAnnTCCC were ranked by Xie et al. as No. 1, 5, 11, 10, and 4, respectively (Xie et al., 2005). Interestingly, I did not discover any novel motif that binds exclusively to bidirectional promoters.

2.3 TRANSFAC motif presence in bidirectional promoters

I continued to use the motif classification framework to compare abundances of known vertebrate motifs from TRANSFAC (Fu and Weng, 2005; Wingender et al., 1996) in bidi-
Figure 2-2: Sequence logos of unique motifs. Motifs discovered by the *ab initio* method are in the second column and corresponding TRANSFAC PSSMs are in the third column.
rectional and unidirectional promoters. Among 604 motifs in the TRANSFAC vertebrate set, 47 are overrepresented, 117 are shared, and 440 are underrepresented. Seven *ab initio* motifs and their corresponding TRANSFAC motifs are consistent in their categorizations, demonstrating the robustness of this approach and the high quality of the position-specific scoring matrices (PSSMs) discovered by MEME. To circumvent the redundancy in TRANSFAC, I calculated Pearson correlations between all PSSM pairs in each category using the malign algorithm (Haverty et al., 2004), setting a cutoff at 0.2, and obtaining 20, 67, and 230 unique motifs for each category, respectively. In both the nonunique and unique cases, the majority (440/604 = 73% and 230/317 = 73%) of TRANSFAC vertebrate motifs are underrepresented in bidirectional promoters (Fig. 2.3). These results are consistent with a model that divergent genes are regulated by a limited set of transcription factors, despite the fact that they make up a substantial portion of human genes.

### 2.4 Positional preference of motifs in bidirectional promoters

Approximately 28% of known motifs show significant positional preference relative to the transcription start sites (TSSs) of human genes (Xie et al., 2005). Because bidirectional promoters have varying lengths, I assessed the positional preferences of the ab initio motifs on normalized promoter lengths. I plot the positional-specific histogram relative to the length of the bidirectional promoter in the column labeled "Relative" (Fig. 2.4), where the number 0 is closest to the TSS of the reverse-strand gene, and the number 1 is closest to the TSS of the forward-strand gene. The positional-specific histograms in absolute bases are also aligned to the reverse-strand gene or the forward-strand gene. The CCAAT box has a bimodal distribution, with peaks at 0.3 and 0.7, corresponding to symmetric placement on the bidirectional promoter around one-third of the way in from each TSS. GABPA has a concave-shaped distribution, peaking close to the flanking TSSs, which is consistent with its position specificity of -23 in unidirectional promoters. Novel 1 appears to have a multipeak distribution. ACTACAnnTCCC, NRF-1, and SP1 display no position specificity, despite showing specificity in unidirectional promoters at -89, -62, and -63, respectively (Xie et al.,
**Figure 2·3: Enrichment of motifs, TF-binding, modified histones, and other nonsequence-specific factors in bidirectional promoters.** The number of binding sites in bidirectional and sampled unidirectional promoters are normalized by the maximum number of sites and plotted against POSSUM log likelihood scores. Counts in the unidirectional reverse gene set are not shown for clarity because they are very similar to the binding-site counts in the unidirectional forward gene set (solid gray). The solid-black lines are the result of summing binding sites in the unidirectional forward and unidirectional reverse gene sets. Error bars indicate the standard error of the mean binding-site counts derived from 10 randomizations. A TF that has a corresponding TRANSFAC motif placed in the same category; a TF whose categorization is consistent among multiple ChIP experiments in different cell types or conditions; a nonsequence-specific factor.

<table>
<thead>
<tr>
<th>Overrepresented</th>
<th>Shared</th>
<th>Underrepresented</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nrf-1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
</tr>
<tr>
<td><strong>SP1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
<tr>
<td><strong>TATA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
<td><img src="image9.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

**Ab Initio Motifs**
- CCAAT
- GABPa
- ACTAGAnnTCCC
- YY1
- Nrf-1

**TRANSFAC Motifs**
- 47 (20 unique)
- 117 (67 unique)
- 440 (230 unique)

**ENCODEx ChIP-chip**
- cMyc
- E2F 1
- E2F 4
- Sp1 (Jurkat)
- Sp3 (Jurkat)
- RNAPII
- TAF
- H3ac
- H3K4me2
- H3K4me3

- cJun
- Sp3 (HCT116)

- STAT1
- Sp1 (K562)
- Sp3 (K562)
- BAF155
- BAF170
- H4ac

---

*Note: (a) corresponds to A TF that has a corresponding TRANSFAC motif placed in the same category; (b) a TF whose categorization is consistent among multiple ChIP experiments in different cell types or conditions; (c) a nonsequence-specific factor.*
One explanation is that the consensus sequences for NRF-1 and SP1 are composed almost entirely of Gs and Cs; thus, their binding sites are frequently found throughout the GC-rich bidirectional promoters. In general, all seven motifs exhibit less positional specificity in bidirectional promoters than in unidirectional promoters, consistent with a model that some of these motif sites are involved in regulating both of the divergent genes.

### 2.5 Divergent genes are co-regulated

Several studies reported that divergent genes have more correlated expression patterns than randomly paired genes (Adachi and Lieber, 2002; Li et al., 2006; Trinklein et al., 2004). As described above, I observed a small set of well-studied motifs to be enriched in bidirectional promoters; thus, I wished to examine whether divergent genes as a group had correlated expression patterns. I computed the Pearson correlation coefficients between the expression profiles (Su et al., 2002) of divergent genes from different bidirectional promoters and indeed observed a significant upward shift in their distribution with respect to that of randomly paired genes from unidirectional promoters (Fig. 2.5; \( P < 2.2 \times 10^{-16} \) by the Wilcoxon rank sum test). Divergent gene pairs from the same bidirectional promoters are even more correlated (\( P = 1.2 \times 10^{-11} \) compared with mismatched divergent genes and \( P < 2.2 \times 10^{-16} \) compared with genes of unidirectional promoters, by the Wilcoxon rank sum test). Setting cutoffs at the fifth and 95th percentiles of the background distribution defined as that of unidirectional promoters, 27% of divergent gene pairs were coregulated, and 1.7% were antiregulated. These figures are close to the 17% and 1.8% I reported earlier, in which I used a different set of microarray data (Trinklein et al., 2004).

### 2.6 TF binding in bidirectional promoters

Taking advantage of publicly available ENCODE ChIP-chip data sets, I analyzed the binding preferences of seven sequence-specific factors, four sequence nonspecific factors, and four types of modified histones in a specified set of regions that make up 1% of the human genome (Birney et al., 2007). I analyzed the target lists of the ChIP-chip experiments
Figure 2-4: Position specificity histograms of *ab initio* motifs in bidirectional promoters normalized by the number of promoters that span each bin. For all the plots, the y-axis represents the number of binding sites with log likelihood score greater than 7, and the x-axis represents the span of the bidirectional promoter either in relative units or absolute base pairs. The second column, labeled Relative, shows the positions relative to both flanking TSSs. The third column, labeled Reverse, shows the position in bp relative to the TSS of the minus strand gene, and the fourth column labeled Forward is relative to the TSS of the plus strand gene. Position specificity (in relative position and bp) of bidirectional promoters and unidirectional promoters are summarized in the fifth and sixth column respectively.
Figure 2.5: The distribution of Pearson correlation coefficients for (A) randomly paired genes of unidirectional promoters, (B) randomly paired divergent genes of different bidirectional promoters, and (C) divergent gene pairs of the same bidirectional promoters. Divergent genes from the same bidirectional promoters are more correlated in expression than randomly paired genes of unidirectional promoters (P = 2.2 x 10^{-16} by the Wilcoxon rank sum test) and randomly paired genes of different bidirectional promoters (P = 1.21 x 10^{-11}).
reported at a 10% false discovery rate cutoff. By comparing the percent overlap (defined as within 500 bp) of these target lists in ENCODE bidirectional promoters with twice the percent overlap in ENCODE unidirectional promoters, I again categorized these factors and modified histones into overrepresented, shared, or underrepresented in bidirectional promoters (see Methods; Fig. 2.3).

The sequence-specific factors I analyzed included JUN, MYC, E2F1, E2F4, SP1, SP3, and STAT1. I found that MYC, E2F1, and E2F4 were overrepresented, as the percent overlap of their target lists with bidirectional promoters was more than twice that with unidirectional promoters at every target-list rank cutoff (Fig. 2.6). Motif analysis above indicates that SP1 and SP3 were shared, but their ChIP data indicates that they preferentially bind to bidirectional promoters in Jurkat cells, show no preference in HCT116 cells, and show preference for unidirectional promoters in K562 cells. Careful examination reveals that this cell-line specificity is slight, consistent with the classification of SP1 and SP3 as shared motifs. In summary, bidirectional promoter binding in living cells for a majority of TFs (E2F1, E2F4, SP1, SP3, and STAT1) are classified in the same way as their canonical sequence motifs.

2.7 Modified histones around bidirectional promoters

Levels of modified histones H3K4me2, H3K4me3, and H3ac have been shown to correlate positively with the level of transcription (Birney et al., 2007). I examined the average levels of these modified histones around individual bidirectional promoters by analyzing ChIP-chip experiments conducted on HeLa cells, and observed that the ChIP intensities tended to be diminished throughout the bidirectional promoter region. However, they were elevated immediately downstream of the flanking TSSs. The average downstream ChIP intensities of all 46 high-quality divergent genes in the ENCODE regions were at similar levels as those of actively transcribed genes and much higher than those of all genes (Fig. 2.7). To summarize the patterns of modified histones, I discretized the ChIP intensities as "elevated" or "diminished" in three regions: downstream of the reverse gene,
Figure 2-6: The fraction of promoters that overlap at least 1 bp with a set of ranked ChIP-chip target lists is plotted on the y-axis. (A) Percent overlap of bidirectional promoters; (B) twice the percent overlap of unidirectional promoters; (C) percent overlap of unidirectional promoters. Both H3K4me3 and MYC overlap more with bidirectional promoters than twice the overlap of unidirectional promoters (see Methods for details).
**Figure 2-7:** Histone modification ChIP intensities downstream of ENCODE genes. Average log2 ChIP intensity of several modified histones and formaldehyde-assisted isolation of regulatory elements (FAIRE) in all genes (Giresi et al., 2007), actively transcribed genes, and 46 divergent genes. Divergent genes have elevated histone modification signals downstream of their TSS, similar to actively transcribed genes.

The most prominent histone modification pattern (32%) consisted of diminished signals in the bidirectional promoter and elevated signal downstream of both TSSs (labeled pattern 1 in Fig. 2.8). Smaller portions had diminished modification signals in the promoter and elevated signal downstream of either the forward TSS (9%; labeled 2 in Fig. 2.8) or the reverse TSS (15%; labeled 3 in Fig. 2.8). These patterns were most pronounced in H3K4me2 and H3ac (Fig. 2.8B,D).

### 2.8 A case study of GABPA

#### 2.8.1 GABPA binds to a majority of bidirectional promoters

I computationally predicted GABPA-binding sites in the human genome and sampled 118 sites randomly from three types of genomic regions for testing: bidirectional promoters, unidirectional promoters, and non-promoters. Binding sites in each region were further
Figure 2.8: Discretized histone modification patterns are not uniformly distributed. The most prominent pattern is labeled 1, where ChIP intensities are elevated downstream of both genes, but diminished throughout the bidirectional promoter. Also prominent are patterns 2 and 3, where the ChIP intensities are just elevated in one downstream gene (the genomic forward direction or reverse direction, respectively). The pattern labeled 4 represents the case where ChIP intensities are elevated in the bidirectional promoter as well as both downstream genes, and 5 represents no elevation throughout the region. The number 6 represents three remaining patterns: elevated ChIP intensities in the bidirectional promoter but diminished in both downstream promoters, elevated ChIP intensities in the bidirectional promoter and downstream of the forward gene, but diminished in the reverse gene, and elevated ChIP intensities in the bidirectional promoter and downstream of the reverse gene, but diminished in the forward gene.
categorized as containing high-scoring motifs, medium-scoring motifs, low-scoring motifs, or no motifs (see Methods for definition of motif scores). Patrick Collins then performed ChIP with an antibody recognizing GABPA and assayed the enrichment of precipitated fragments by quantitative real-time PCR (qPCR) in Jurkat and K562 cells. Figure 2.9 shows the breakdown of GABPA-bound fragments in the four motif categories at the five-fold enrichment cutoff. These results show that GABPA binds to the majority (83%) of tested bidirectional promoters with high-scoring motifs (compared with 51% in unidirectional promoters and 14% in nonpromoters). As 21% of bidirectional promoters have at least one high-scoring motif, I estimate that 17% (= 83%*21%) of bidirectional promoters have high-scoring motifs AND are bound by GABPA. Adding this to the estimates for medium- (16%) and low- (24%) scoring motifs gives a total of 57% of bidirectional promoters bound by GABPA. In contrast, it is estimated that only 7% of unidirectional promoters are bound by GABPA. This genome-wide estimate agrees well with the results in the ENCODE regions. Patrick Collins tested 16 of the 23 bidirectional promoters and found that nine (56%) were bound by GABPA in two cell lines and 12 (75%) were bound by GABPA in at least one cell line. Thus, GABPA can bind to a majority of bidirectional promoters, and its binding frequency is overrepresented in bidirectional promoters compared with unidirectional promoters. This suggests that GABPA is a major regulator of bidirectional transcription. A follow up study with 241 additional promoters showed that GABPA bound to 98/113 (87%), 91/115 (79%), and 73/117 (62%) promoters in Jurkat, K562, and HeLa cells. By comparison, 58/234 (25%), 67/230 (29%), and 50/275 (18%) of general promoters were bound in the same cell types (Collins et al., 2007).

2.8.2 GABPA motif presence correlates with binding in living cells

The higher the motif score is for a site, the higher the likelihood that the site is occupied by GABPA. In bidirectional promoters, 44% of low-scoring sites were bound by GABPA compared with 68% of medium-scoring and 83% of high-scoring sites. Similarly, in unidirectional promoters, 0% low-, 42% medium-, and 51% high-scoring sites were bound by
Figure 2.9: GABPA ChIP in three region categories and four motif score cutoffs. GABPA binds preferentially to bidirectional promoters over unidirectional promoters and over nonpromoter regions at all motif score cutoffs. Regions containing high-scoring binding sites are more likely to bind GABPA than those containing only medium or low-scoring sites. NA indicates that no regions in this category were tested.
GABPA. All regions tested without sites showed no binding, suggesting that the GABPA consensus sequence is necessary for binding in living cells. This appears to be unique to GABPA; other sequence-specific factors have been reported to bind many regions that do not contain a match for their canonical sequence motif (Bieda et al., 2006). GABPA could be driven to bind bidirectional promoters because a larger proportion of bidirectional promoters contain medium- and high-scoring binding sites (Fig. 2.10A), and the number of binding sites is also greater (Fig. 2.10B) compared with unidirectional promoters.

Eighty-six percent of the fragments Patrick Collins tested by ChIP contained more than one GABPA site. The fragments with the greatest enrichment usually contained one high-scoring site surrounded by several low-scoring sites. I sought to determine which characteristic - the score of the best binding site or the total number of binding sites - better predicted ChIP enrichment. I set the score threshold to the low level (log-likelihood score 3) and counted the total number of sites in the ChIP-identified fragment. The largest number of sites was seven, in a 185-bp long bidirectional promoter, while the average was three. The correlation between the total site count and ChIP enrichment (correlation coefficient = 0.40) is stronger than that between the highest site score and ChIP enrichment (correlation coefficient = 0.29), suggesting that having multiple sites is a better predictor of binding than having a single high-scoring GABPA site. In fact, these two characteristics frequently coincide; thus, it is possible that low-scoring sites help guide sequence-specific factors to high-scoring sites.

2.8.3 GABPA-binding sites often required for bidirectional activity

In a previous study, Nathan Trinklein systematically truncated and tested nine bidirectional promoters for activity in both directions by using a luciferase reporter assay (Trinklein et al., 2004). In the current study, I reanalyzed the truncated fragments by mapping binding sites of the aforementioned ab initio motifs in these nine promoters and found that eight of them contain at least one GABPA site that appears to be necessary for their bidirectional transcriptional activity. I define a binding site as necessary if the promoter
Figure 2-10: GABPA binding site presence by proportion (A) and by raw frequency (B). (A) Percent of bidirectional (1,304) and unidirectional (13,205) promoters containing at least one GABPA binding site at various motif score cutoffs. Recall high scoring sites have score 9 or higher, medium scoring sites are between 8 and 9, while low scoring sites are between 3 and 8. A larger proportion of bidirectional promoters contain medium and high scoring sites, whereas a larger proportion of unidirectional promoters contain low scoring sites. (B) Binding site frequencies normalized by the maximum count in the plot. Bidirectional promoters contain more GABPA binding sites with score greater than 6.
activity was diminished in both directions after deletion of a fragment containing that site. The large number of necessary GABPA sites also supports its role as a major regulator of bidirectional promoters. However, in six of nine cases deletions of portions of the bidirectional promoter that did not have a GABPA site also showed diminished activity, indicating that there are other factors necessary for bidirectional promoter activity. For the case in Figure 2.11, a 30-bp region containing a single GABPA site was sufficient to drive transcription in both directions. Mutagenesis results further confirmed that 12 bases beginning at position 7 and overlapping the GABPA consensus CCGGAARYR are essential for bidirectional promoter activity (Fig. 2.12). The footprint of GABPA may extend beyond the 9-bp consensus sequence, as an additional three bases were found to be important for transcription (Mutation 9 in Fig. 2.12). Mutations 14 and 1 also led to decreased promoter activity; however, they are at the ends of the fragment and the decrease in activity occurs in only one direction; thus, they most likely correspond to the transcription start sites of the forward and reverse genes, respectively.
### Figure 2.12: Mutagenesis of a 30-bp bidirectional fragment.

Bases essential for promoter activity in each direction are underlined. Transversion mutations were introduced three bases at a time and shifted two bases for a total of 15 mutants. Mutation of a 12-bp region annihilates promoter activity in both directions. The consensus CCGGAARYR is recognized by the transcription factor GABPA. Asterisks (*) indicate position of transcription start sites.

### 2.9 Methods

#### 2.9.1 Two bidirectional promoter data sets

The first set of bidirectional promoters was selected to maximize quality and used for the *ab initio* motif search. TSS definitions were taken from DBTSS version 4.0 (Yamashita et al., 2006), and the 500-bp or shorter intergenic regions between TSSs of oppositely stranded nonoverlapping transcripts were taken to be bidirectional promoters. This results in 376 bidirectional promoters.

The second set of bidirectional promoters was selected to increase quantity and used for binding-site enumeration and motif categorization. The 1-kb or shorter intergenic regions between oppositely stranded nonoverlapping transcription units (TUs) were taken to be bidirectional promoters. TUs were constructed by combining TSS annotations from three databases in order of preference: (1) DBTSS (Yamashita et al., 2006), (2) hg16 RefGene table from UCSC (Karolchik et al., 2003), and (3) PromoSer (Halees and Weng, 2004). For those transcripts that were neither in DBTSS nor RefSeq, I used the 5'-most PromoSer TSS mappings with quality >2, which correspond to TSSs based on GenBank.
mRNAs excluding ESTs. Each transcript was then consolidated into TUs defined by PromoSer cluster identifiers. A total of 32,349 TUs were identified, and of the following composition: 9767 derived from DBTSS, 5228 derived from RefGene, and 17,354 derived from PromoSer. This results in 1304 bidirectional promoters. I believe that this larger bidirectional promoter set balances quality and quantity in the selection process.

2.9.2 One unidirectional promoter data set

Using the previously described transcription units, I annotated 13,205 unidirectional promoters whose closest TU neighbor is >1 kb away. A total of 6503 unidirectional promoters are oriented in the forward genomic direction, while 6702 are oriented in the reverse genomic direction.

2.9.3 Gene ontology analysis

The names of divergent gene pairs from the second set of bidirectional promoters were extracted and analyzed using GO::TermFinder (Boyle et al., 2004). GO::TermFinder analyzes a list of genes to determine whether any GO terms occur more frequently than would be expected by chance. The P-value was calculated using the hypergeometric distribution and the Bonferroni correction was applied to correct for multiple testing. I set a corrected P-value cutoff of 0.1 and obtained a limited list of enriched GO terms.

2.9.4 Motif discovery

Sequences from the first bidirectional promoter data set were extracted and analyzed using MEME Version 3.5.0 (Grundy et al., 1996). I requested output of 15 motifs and further narrowed this list down to seven unique motifs by pairwise correlation of the 15 motifs (Haverty et al., 2004). Two motifs were poly(A) and poly(T) repeats, and six motifs were reverse complements of other motifs in the set.
2.9.5 Motif representation calculation

Along with the *ab initio* discovered motifs, a set of known vertebrate PSSMs were extracted from TRANSFAC (Fu and Weng, 2005; Wingender et al., 1996), and mapped onto the large bidirectional promoter set using POSSUM (http://zlab.bu.edu/~mfrith/possum/). POSSUM calculates the log likelihood ratio of observing a subsequence given the motif definition versus observing the subsequence given the nucleotide composition in a 100-bp window around the subsequence. For background comparison, I individually checked the CpG dinucleotide content of every bidirectional promoter and sampled without replacement for a corresponding unidirectional promoter of the top 10 unidirectional promoters with closest CpG dinucleotide content. This results in an average CpG dinucleotide content of 7.42% in random samples compared with 7.28% in all bidirectional promoters, and a C+G content of 61.83% in random samples compared with 61.54% in all bidirectional promoters. The binding sites for all of the motifs were enumerated at different POSSUM log likelihood score cutoffs in all bidirectional promoters as well as the random samples. I averaged the classifications obtained by two methods to place motifs in the overrepresented, shared, or underrepresented category. The first method by voting determines whether the motif counts in bidirectional promoters (dotted line in Fig. 2.3) are predominantly (1) above the unidirectional sum (solid black line), (2) between the unidirectional sum and the single unidirectional counts (solid black and solid gray line, respectively), or (3) below the single unidirectional counts (solid gray line). The second method by distance calculation determines whether the motif counts in bidirectional promoters are closer to (1) the unidirectional sum, (2) the single unidirectional counts, or (3) the horizontal line with a height of 0, by summing the Euclidian distances between the motif counts at each log likelihood cutoff. Overrepresented, shared, and underrepresented motifs were coded as 1, 2, and 3, respectively, and the codes from the two methods were averaged across 10 randomized picks of unidirectional promoter sets. Overrepresented motifs have an average code $\geq 2.5$, shared motifs have an average code $< 2.5$ and $\geq 2$, and underrepresented motifs have an average code $< 2$. 


2.9.6 ENCODE ChIP-chip factors in bidirectional promoters

Genomic target lists for various ChIP-chip experiments were reported at 10% false discovery rate cutoff by the ENCODE Consortium. On average, each target was 765 bp in length, although the number of targets varied dramatically depending on the factor. Because the scope of the ChIP-chip experiments was within the ENCODE regions (1% of human genome), I extracted bidirectional and unidirectional promoters in these regions using previously defined TUs that overlap with the ENCODE regions, resulting in 23 bidirectional promoters (46 divergent genes), and 227 unidirectional promoters. A target is said to overlap a promoter (bidirectional or unidirectional) if at least one base pair of the target overlaps within a 500-bp window around the promoter; hence, a portion of the downstream region is also captured in the overlap analysis. This is important, as histones are depleted at the promoter but not downstream of the TSS. I report the percentage of bidirectional promoters (and unidirectional promoters) that overlap with the target list at every rank cutoff. Factors that are overrepresented show greater overlap in bidirectional promoters than twice the overlap in unidirectional promoters across 70% of the rank cutoffs (the solid line labeled A in Fig. 2.6 is above the dash-dot line labeled B 70% of the time).

2.9.7 Histone modification analysis

I analyzed raw ChIP-chip enrichment scores of four modified histones (H3ac, H4ac, H3K4me2, and H3K4me3), all in HeLa cells, after 30 min of retinoic acid stimulation. The results do not change if I separately analyze or include ChIP signal intensities of modified histones before stimulation. To categorize histone-modification patterns, I averaged the overall signal intensities in the diminished intergenic region and the elevated downstream 1-kb region, and checked whether the average signals of three regions in individual bidirectional promoters (1 kb downstream of the reverse gene, bidirectional promoter region, and 1 kb downstream of forward gene) were closer to the overall elevated or diminished averages. For example, the average H3ac signal intensity is 1.90, 1 kb downstream of the reverse strand gene NM_024298, and 1.29, 1 kb downstream of the forward strand gene NM_024075. The
bidirectional promoter regulating both genes has an average signal intensity of 0.80. Comparing each number from these three regions to the elevated (1.54), and diminished (0.91) H3ac signals of all divergent genes, I categorize the first and second region as elevated, and the third region as diminished, because the first and second numbers are closer to 1.54, while the third number is closer to 0.91. The ChIP intensities in the overall diminished intergenic region are typically half that of the overall elevated downstream 1-kb regions.

2.9.8 Microarray analysis

I used binary (MAS5 absent/present) expression profiles from Affymetrix U133 array to analyze the overall expression levels of all genes and coexpression levels of divergent genes in the ENCODE regions (Koch et al., 2007). To correlate gene-expression profiles, I used the GNF data set of 156 Affymetrix U133 microarray experiments across 78 human cell types (Su et al., 2002). Because a gene can map to several probesets on the microarray, I computed the mean of all pairwise correlation coefficients between probeset pairs. I did not use the signed absolute maximum as in an earlier study (Li et al., 2006), because this systematically exaggerated the correlations and resulted in bimodal distributions for all gene sets examined - even the randomly paired genes of unidirectional promoters. The earlier study used randomly paired probesets as the background, and hence, did not reveal this artifact (Li et al., 2006).

2.9.9 Deletion constructs

Binding sites of the ab initio discovered motifs were mapped to eight deletion construct experiments using the cis-element prediction program POSSUM (score cutoff = 7), and visualization program MotifViz (Fu et al., 2004). Deletion construct functional assays were previously conducted and reported by Trinkelin et al. (Trinklein et al., 2004).

2.9.10 Mutation analysis of 30-bp fragment

Nathan Trinklein generated 3-bp substitution mutations throughout a 30-bp region in one bidirectional promoter and tested each for promoter activity in both directions. He pro-
duced these mutations by using random transversions (Pu → Py; Py → Pu) three bases at a time, and shifting by two bases. Both strands of each mutant construct were synthesized by Operon (standard phosphoramidite synthesis), annealed, cloned, and sequence verified. The transcriptional activity was assayed in HT1080 cells in two directions using the luciferase reporter vector described in (Trinklein et al., 2004).

2.9.11 GABPA ChIP-qPCR

Genomic GABPA-binding sites were computationally predicted using POSSUM, and 118 sites stratified by region (bidirectional promoter, unidirectional promoter, or nonpromoter), and binding site score (high, medium, low, or no motif) were randomly selected for testing. Binding sites are called high scoring if they have a POSSUM log likelihood score $>9$; medium scoring if they have a score between 8 and 9; and low scoring if they have a score between 3 and 8.

Sonicated chromatin from either $2 \times 10^7$ K562 cells or $4 \times 10^7$ Jurkat cells was incubated for 24 h with 5 g of mouse monoclonal GABPA antibody (catalog # sc-28312, Santa Cruz Biotechnology) coupled to sheep anti-mouse IgG magnetic beads (Dynal Biotech). The magnetic beads were washed five times with buffer containing 100 mM Tris, 500 mM LiCl, 1% NP-40, and 1% deoxycholate, and once with TE buffer (10 mM Tris at pH 8.0, 1 mM EDTA). The DNA was then eluted by incubating in buffer containing 1% SDS and 0.1 M NaHCO3 at 65 for 12 h. After removing the magnetic beads, the eluent was further incubated for 16 h at 65 to reverse the cross-links. A phenol chloroform extraction was performed and the aqueous phase desalted and concentrated using the QIAquick PCR purification kit (QIAGEN).

Patrick Collins performed real-time PCR to measure the enrichment at each promoter by incorporation of SYBR-green. He designed primers around the highest-scoring GABPA motif in a promoter, so that the final amplicon was 60-100 bp in length. Reactions were performed according to Bio-Rad recommendations for detection on the iCycler instrument. For each amplicon, he constructed a standard curve of threshold cycles from 50 ng, 5 ng, 500
pg, and 50 pg of genomic DNA (Roche). He then fit the threshold cycle of ChIP-enriched DNA to determine the quantity of starting template. To determine the fold enrichment for any particular fragment, the quantity of starting template was divided by the average starting quantities of three negative controls.
Chapter 3

Tissue-specific Transcription Regulation

3.1 Introduction

Gene regulation gives rise to diverse morphologies, functions, and dynamic responses of the approximately 210 cell types in the human body (Su et al., 2002; Kadonaga, 2004; Levine and Tjian, 2003; Mitchell and Tjian, 1989). In mammals, sequence elements and nuclear proteins, including core promoter elements, enhancers, repressors, and epigenetic modifications, interact to drive expression of genes. These interactions have been revealed through a large variety of methodologies, including global studies of transcripts (Carninci et al., 2005; Harrow et al., 2006; Wakaguri et al., 2008), binding of proteins to DNA (Birney et al., 2007; Johnson et al., 2007), enhancer reporters (Pennacchio et al., 2006), DNA methylation studies (Weber et al., 2007), among others (Birney et al., 2007; Cooper et al., 2006; Johnson et al., 2007; Trinklein et al., 2003). Among the variety of elements that drive gene expression, the promoter was among the first and best characterized. A promoter is functionally defined as the core sequence elements necessary and sufficient to drive transcription. Whereas the size and nature of promoters varies widely, core sequence elements are usually contained in a short segment of the genome, from approximately 1kb upstream to approximately 100bp downstream of the transcription start site (TSS). Core promoter elements may drive ubiquitous expression, or they may drive expression only in a particular biological condition. Encoded within a promoter are myriad sequence features, or motifs, to which transcription factors bind and drive function (Birney et al., 2007; Brown et al., 2007; Cooper et al., 2006; Johnson et al., 2005; Johnson et al., 2007; Lin et al., 2007; Myers et al., 1986). The identification and detailed characterization of these motifs as well
as the interactions among them would be an important step forward in understanding the fundamental mechanisms of gene regulation.

It has proven difficult to identify and characterize human promoters on a large scale and with high resolution. A variety of methods have been proposed, including alignment of full length cDNAs (Imanishi et al., 2004), mapping of 5' ends by CAGE tags (Shiraki et al., 2003), chromatin immunoprecipitation (ChIP) on arrays (Birney et al., 2007; Kim et al., 2005b; Kim et al., 2005a), transient transfection assays (Cooper et al., 2006; Myers et al., 1986; Trinklein et al., 2003), and a combination of methods (Kim et al., 2005b). Expression based methods (CAGE tags, cDNA, expression microarrays) directly measure transcript abundance in tissues, and can be performed in a highly parallel fashion. While the promoter sequence can be assumed to be upstream of the 5' end of a CAGE tag or cDNA, it is difficult to specifically isolate or discover new promoter sequences using expression microarrays. On the other hand, expression-based methods are often among the most efficacious methods to characterize gene expression in vivo, as they capture the complex effects of epigenetic interactions, long-range enhancers, and other genomic features that are difficult to recreate experimentally.

Among the most direct methods for characterization of functional promoter sequences is the transient transfection promoter activity assay. In this classic method (Cooper et al., 2006; Myers et al., 1986; Trinklein et al., 2003), a putative promoter sequence is fused in a plasmid construct with a reporter gene, such as luciferase, and the recombinant construct is introduced experimentally into mammalian tissue culture cells. This method is advantageous in that it directly measures the function of a specified DNA fragment, and isolates the core sequence features necessary and sufficient to induce transcription. However, this method has the disadvantage of removing the promoter from its native genomic context, such that enhancers, epigenetic modifications, and other contextual factors are removed. Regardless of this caveat, transient transfection reporter assays often yield tissue-specific information (Cooper et al., 2006). For example, the 5’ proximal promoter region of the hepatocyte growth factor (MET) gene was most highly active in the liver-derived cell line
HepG2 (Cooper et al., 2006). The osteoclast-associated receptor (OSCAR) 5’ proximal promoter was active in only 4/16 of the cell lines tested, one of which is mg63, the osteosarcoma cell line (Cooper et al., 2006). The promoter activity assay still remains the gold standard for promoter identification and characterization.

To better characterize human promoter function globally, I collaborated with Richard Myers from Stanford University, building upon previous studies (Cooper et al., 2006; Trinklein et al., 2003), to implement technologies for high-throughput promoter activity assays. David Johnson, the senior scientist in the Myers lab, devised a technology that increases the throughput of these assays more than fourfold above the previous technology, and tested 4,575 putative human promoters across eight immortal cell lines. In parallel, he performed genome-wide gene expression analysis for each of these cell lines. This is by far the largest promoter activity data set ever produced, and the only comprehensive promoter transient transfection assay data set that includes genome-wide gene expression analysis in the corresponding cell types. This represents an unprecedented opportunity to explain endogenous expression by deconvoluting the contribution of core promoter elements from the contribution of epigenetic and long range elements. Preliminary analysis of HpaII cut sites in HCT116 cells determined by ChIP-seq indicates that DNA methylation information can be combined with promoter activity to explain a larger proportion of the variance observed in endogenous expression. By combining complementary data sets in matched cell lines, these types of meta-models look broadly at endogenous expression, constructing a systematic framework for measurable components, and quantifying the variance one can not yet explain.

To take a more in depth look, I identified sequence features that drive human promoter function and tissue-specificity by modeling the effects of 691 TF motifs both simultaneously and individually. As I was interested in TFs that contribute to promoter activity, the modeling strategy is formulated such that TFs are required to impact the expression of a downstream gene. Half of the TFs identified in the computational screen recapitulated factors known to function in the corresponding cell type and the remaining half were novel
predictions. For many of the cell lines tested, I found motifs that were associated with expression specific to only one cell line. This work provides a comprehensive characterization of tissue-specific TF motifs in human promoters, and may be used to prioritize TFs for further study.

### 3.2 Promoter activity and gene expression data

The first step in the process of identifying and characterizing human promoter sequences is the *in silico* prediction of putative transcription start sites (TSSs). For this step, Nathan Trinklein, from SwitchGear Genomics, made use of existing data that describe human transcribed sequences, such as full-length cDNAs and 5’-end cDNA sequence tags. He computationally aligned a database of over 250,000 human cDNAs to the reference human genome, and predicted 37,000 gene models, with 22,000 gene models represented by two or more cDNAs. He then used 5’-end tags (Carninci et al., 2005; Ruan et al., 2007; Wakaguri et al., 2008) to formulate a score for the TSS of each gene model. The TSS score assigns a confidence level to each promoter prediction, with the largest bin of scores below 20. Generally, lower scores produce fewer positives, but potentially contain more novel, uncharacterized promoter types. David Johnson selected 4,575 of these TSS predictions (2,083 had TSS scores below 20), and built promoter transient transfection constructs. The promoter constructs cover nearly all of the putative promoters on chromosome 7, as well as 2,266 promoters from various genomic loci across the genome. They contained on average 1kb of sequence spanning the putative promoter, from 900bp upstream of the TSS to 100bp downstream of the TSS. The Stanford team then conducted transient transfections for each of these 4,575 promoter constructs in eight immortal cell lines (Ht1080, G402, T98g, Hct116, HeLa, HepG2, Ags, and U87mg). They specifically chose these cell lines to represent a variety of parent tissue types, from hepatocyte (HepG2) to neuroblastoma (t98g). Across three replicates, 4,575 constructs, and eight cell lines, they performed 109,800 transient transfection experiments.

As a complement to the transient transfection data and as a molecular phenotype of
the cell lines, David Johnson also conducted genome-wide gene expression analysis using Human RefSeq 8 BeadArrays from Illumina. Though they are not direct measurements of promoter activity, one can deduce certain types of information from the expression data. Specifically, there is a subset of 1333 transcripts measured by the expression arrays, which can be unambiguously matched to a single putative TSS. These genes lack putative alternative promoters, and therefore the matched promoter regions can be used as a complementary data set to the corresponding gene expression data. In total, the endogenous expression of 20,589 genes were characterized in 8 cell lines.

The distributions of promoter activity and gene expression scores was bimodal, indicating that both data sets likely arise from a population of inactive promoters/untranscribed genes, and another population of active promoters/transcribed genes. This allowed me to choose a reasonable threshold for all 8 cell-lines at the trough of the distribution to separate the two peaks (vertical dashed line in Fig. 3.1A). The promoter activity threshold was defined at log 2 promoter activity score of 0, corresponding to promoter activity score of 1 - the point at which the luciferase signal exceeds the background signal. A threshold for endogeneous expression was likewise defined at the trough of the log 2 gene expression score distribution, corresponding to log 2 gene expression score of 7 (vertical dashed line in Fig. 3.1B). Upon this thresholding, 67.0% of putative promoters were active in at least one cell line, and 54.4% of genes were expressed in at least one cell line.

To look across cell lines, promoter activity and endogenous expression profiles were encoded as bit-strings of length 8, with 1 indicating that a promoter/gene is active/transcribed in a given cell line and 0 otherwise. The frequencies of the 20 most common promoter activity profiles are plotted in Figure 3.2. Most promoters and genes showed ubiquitous activity and expression. Among the 4575 putative promoters, 789 (17.2%) of putative promoters were active in all 8 cell lines, 483 (10.6%) were active in only one cell line, and 1508 (33.0%) were inactive in all tested cell lines. Among 20,589 genes, 25.3% (5209) were expressed in all 8 cell lines, 7.9% (1619) were transcribed in only one cell line, and 45.6% (9393) were unexpressed in all tested cell lines.
Figure 3.1: Distribution of Promoter Activity and Gene Expression Scores. Thresholds for active/inactive promoters and expressed/unexpressed genes were defined at the troughs of bimodal distributions, indicated by the vertical dashed lines.

3.3 Genetic and epigenetic contributions to endogenous expression

Interestingly, the proportion of genes that were ubiquitously expressed (25.3%) or unexpressed (45.6%) across eight cell lines (71%) was larger than the proportion of promoters that were ubiquitously active (17.2%) or inactive (33.0%) across eight cell lines (50%). I ruled out the possibility that alternative promoters tested in the transient transfection assay might have contributed to more nuanced promoter activities as they can be regulated in a tissue-specific manner (Kamat et al., 2002). When I repeated the analysis in 1333 matched promoters and genes that lacked alternative TSS, the same trend was observed where a larger proportion of genes were ubiquitously expressed or unexpressed (72.9%) while a smaller proportion of promoters were ubiquitous active or inactive (47.4%). I also examined the proportion of ubiquitous promoters and genes at increasingly stringent cutoffs (25th, 65th, 75th, 90th quantiles), to rule out the possibility that the observed trend was due to thresholding artifacts. Indeed, at every gene expression and promoter activity
Figure 3-2: Twenty most abundant binarized promoter activity patterns. The number ‘1’ indicates that a promoter is active in a given cell line and the number ‘0’ indicates otherwise.
threshold combination, the proportion of ubiquitous genes exceeded the proportion of ubiquitous promoters (Fig. 3.3). This result suggest a model where the default state of core promoters is more nuanced compared to endogenous expression. Core promoter elements may drive varied tissue-specific activities, whereas epigenetic, chromatin, and long range elements create more uniform endogenous gene expression.

On average, the Pearson correlation coefficient between the promoter activity and endogenous expression of 1333 matched promoters and genes was 0.43 compared to $6.5 \times 10^{-4}$ for randomly matched promoters and genes. This is in close correspondence to a previous study, where the correlation between promoter activity and endogenous RNA transcript levels (as determined by quantitative RT-PCR) was reported to be 0.53 (Cooper et al., 2006), indicating that core promoter elements and sequence features can explain approximately 19% to 28% of endogenous expression levels overall ($p < 2.2 \times 10^{-16}$). I investigated the contribution of DNA methylation by overlapping ChIP-seq target lists for HpaII cut sites with the 1333 matched promoters and genes. Overall, the Pearson correlation coefficient between endogenous expression and promoter activity was 0.38, while the overall correlation between endogenous expression and DNA methylation is 0.32. The partial correlation between endogenous expression and DNA methylation controlling for promoter activity is 0.23, indicating that DNA methylation can additionally explain ($0.23^2 = 5\%$) of the variance seen in endogenous expression. A meta-model incorporating core promoter activities and DNA methylation to explain endogenous expression in HCT116 cells yields an $R^2$ of 0.19 ($p < 2.2 \times 10^{-16}$).

### 3.4 Novel promoters

Each putative promoter was given a TSS confidence score based on transcript evidence (e.g. human cDNAs, 5’ end tags) and the TSS confidence score was a strong predictor of average promoter activity (Pearson correlation coefficient $r = 53\%$). I intentionally selected 2083 putative promoters with low TSS scores ($< 20$) to discover and characterize potentially novel promoters. In total, these screens revealed 1082 novel promoters that
Figure 3-3: Fraction of ubiquitous promoters and genes. Increasingly stringent thresholds for calling active promoters are defined along the rows at the 58th, 75th, and 90th quantiles of the promoter activity score distribution. Similarly, increasingly stringent thresholds for calling expressed genes are defined along the rows at the 61st, 75th, and 90th quantiles of the endogenous expression score distribution. At every threshold combination, the proportion of ubiquitously expressed and unexpressed genes is larger than the proportion of ubiquitously active and inactive promoters.
Figure 3.4: Fraction of novel and known active promoters. Each bar represents the fraction of novel (light gray) or known (dark gray) promoters that are active in 1-8 cell lines (x-axis). Novel promoters are significantly active in fewer cell lines than known promoters for both LCG (p < 2.2 x 10^{-16}) and HCG promoters (p = 1.847 x 10^{-5} by Mann-Whitney non-parametric test).

were active in at least one cell line, contributing approximately 30% more promoters to the current repertoire of human promoters (3067) in 5% of the human genome. A larger proportion of novel promoters had low CpG content, 66.8% (723/1082), compared with only 24.7% (491/1985) in known promoters. Novel promoters tended to be active in fewer cell lines, with 26.7% (289/1082) being specifically active in only one cell line. This trend is illustrated in Figure 3.4, and was most pronounced among novel LCG promoters (p < 2.2 x 10^{-16}), and also significant among novel HCG promoters (p = 1.847 x 10^{-5}). As 26.7% of novel promoters were cell line-specific, many will serve as important training examples in the models, garnering the final set of motif modules that are predicted to regulate cell linespecific promoter activity. Identifying these novel promoters is an important step toward understanding unique mechanisms of tissue-specific transcription regulation.
3.5 Predictive models of promoter activity

In this section, I describe a modeling strategy to identify sequence features that drive human promoter function and tissue-specificity. I modeled the effects of 691 TF motifs both simultaneously and individually, and the modeling strategy was formulated such that TFs are required to impact the expression of a downstream gene. Before any models could be implemented, I explored two important aspects of the modeling strategy in detail: How to represent the sequence features in a promoter, and which machine learning algorithm to use in the models.

3.5.1 Exploratory analysis of motif representation

From the study of GABPA described in Chapter 1, I learned that the ab initio-discovered GABPA position specific scoring matrix was highly predictive of in vivo binding. The higher the motif score for a site, the higher the likelihood that the site was occupied by GABPA. Additionally, because certain genomic regions tend to be in open chromatin states, the type of genomic region highly influenced in vivo binding, with GABPA binding to the majority (83%) of bidirectional promoters with high-scoring motifs compared with 51% in unidirectional promoters and 14% in non-promoters. Fortunately, the current study is only looking at promoters, which limits TF predictions to regions of open chromatin, increasing overall accuracy.

Interestingly, the GABPA ChIP fragments with the greatest enrichment usually contained one high-scoring site surrounded by several low-scoring sites. The correlation between the total site count and ChIP enrichment (correlation coefficient = 0.40) is stronger than that between the highest site score and ChIP enrichment (correlation coefficient = 0.29), suggesting that having multiple sites is a better predictor of binding than having a single high-scoring GABPA site (Fig. 3.5). To incorporate the score of the strongest TF binding site as well as the abundance of low-scoring sites surrounding it, I decided to use the algorithm Clover (Frith et al., 2004a) to obtain a composite score for each motif in each promoter, quantifying the equilibrium binding potential of a TF to a promoter.
according to a thermodynamic model (Stormo, 2000). Motifs were represented by position specific scoring matrices (PSSMs), which were collected from three sources: the TRANSFAC database (Matys et al., 2006), previous analyses in the human genome (Lin et al., 2007; Wei et al., 2006; Xi et al., 2007), and a catalogue of mammalian motifs (Xie et al., 2005).

3.5.2 Exploratory analysis of modeling strategies

I investigated whether a statistical model built solely on sequence motif information could predict the promoter activity status (active vs. inactive) of the 642 assayed promoters in the ENCODE regions. First, I divided the promoters into two classes, and built separate models for the High-CpG and Low-CpG classes. These two classes naturally separate in the human genome and may employ different regulatory mechanisms by different sets of TFs (Saxonov et al., 2006). For example, housekeeping genes tend to be regulated by high CpG content promoters that lack a TATA-box, while tissue specific genes tend to be regulated by low CpG content promoters that contain a TATA-box. I then computed Clover raw scores (Frith et al., 2004a) of 910 motifs from the TRANSFAC database and Xie et al. (Xie et al., 2005) to quantify the TF profile on each promoter. This was most appropriate because Clover raw scores incorporate high scoring single TFBS, as well as lower scoring TFBS along the entire promoter, thus taking into account the genomic context of the motif. These scores were used as predictors in six machine learning algorithms (gradient boosting machine, linear discriminant analysis, random forests, neural network with five internal nodes, neural network with six internal nodes, and support vector machine), whose task was to discriminate between active and inactive promoters. The 75th percentile normalized promoter activity score among 16 cell lines needed to be greater than 2, in order to be called an active promoter. The baseline accuracy was 50%, as 318 out of 642 promoters were actively transcribed (black dashed line in Fig. 3.6). Similarly, 68% (204/300) of High-CpG promoters were active, while 33% (114/342) of Low-CpG promoters were active. All six algorithms performed significantly better than the baseline (Fig. 3.6). Linear discriminant
Figure 3.5: Linear relationship between GABPA sites and ChIP fold enrichment. GABPA ChIP was performed on 118 genomic regions in Jurkat and K562 cell lines.
analysis performed the worst, suggesting that some of the interactions among TFs could be non-linear. I decided to use SVMs for this project because it performed well and is very versatile.

In order to prioritize a subset of TFs for experimental testing, I will narrow down a list of 691 motifs by their relative importance in predicting promoter activity. I anticipate that the final list will consist of approximately 50 TFs. A naïve feature selection task was performed to examine the baseline level of feature reduction that can be achieved. For each learning task, I performed a t-test on each motif for its ability to discriminate active from inactive promoters. I kept a tally of the significantly influential motifs based on a Bonferroni adjusted p-value cutoff of 5%, and eliminated motifs that were never picked
as predictors because their adjusted p-value was greater than 5% in all 20 trial runs. This strategy yielded 451 motifs, which was further reduced to 301 motifs by eliminating redundancy among the TRANSFAC PSSMs.

An important result from this analysis was that the motifs useful for discriminating active from inactive promoters in the High-CpG class did not overlap with those useful in the Low-CpG class. Additionally, these motifs would not be differentiated if both classes were analyzed together. For example, the most important known motifs in the High-CpG class included E2F1, Pax, ZF5, ETF, Nrf-1, Whn, AP-2, and NF-Y, while the most important motifs in the Low-CpG class included TATA, NERF1a, Nkx2-2, POU3F2, IPF1, NKKx6-1, FOXJ2, Tst-1, and the CACCC-binding factor. I expected NF-Y (recognizes CCAAT-box) to be important in the High-CpG class because it is known to bind many TATA-less and high CpG promoters (Mantovani, 1999). Similarly, I expected TATA to be important in the Low-CpG class because TATA-containing promoters tend to have low CpG content. It is promising that a naïve feature selection strategy can reduce the motif set by 1/3, and produce biologically meaningful stratifications. More sophisticated feature selection algorithms can most likely reduce the list of influential motifs further.

### 3.6 The effect of CG dinucleotide

In order to predict promoter activities, I used CG dinucleotide content as a feature in the SVM model in addition to the catalogue of 691 vertebrate motifs. The CG dinucleotide counts were divided by promoter length, and were remarkably predictive of ubiquitous promoter activity, with $r=0.75$ and AUC=94% (area under the Receiver Operating Characteristic (ROC) curve), surpassing that of any single motif. Note that AUC=100% represents the ideal discriminator and AUC=50% represents a random discriminator. Because the CG dinucleotide is embedded in many general TF motifs, for example Nrf-1 (CGCATGCGC) and E2F1 (TTGCAGGC), one explanation for the high predictive ability of CG dinucleotide content is that it combines the effect of all the CG-rich motifs in a promoter, making it difficult to tease out the contributions of individual TFs.
To overcome this effect, I grouped the promoters by CG dinucleotide abundance. The normalized CG dinucleotide content (defined in Methods) of the 4,575 promoters displayed a bimodal distribution (Fig. 3.7), similar to that of all human promoters reported previously (Saxonov et al., 2006). I define Low-CG promoters (LCG) as having normalized CpG < 0.5, whereas High-CG promoters (HCG) have normalized CG $\geq$0.5 (Fig 3.7). It is known that HCG promoters tend to be ubiquitously active, whereas LCG promoters tend to be tissue-specific (Saxonov et al. 2006). Indeed, promoters that were active in all eight cell lines were dominated by HCG promoters (91%; 719/789), whereas only 9% (70/789) were LCG promoters. Among cell-line specific promoters that are active in only one cell line, only 21.7% (105/483) are HCG whereas 78.3% (378/483) are LCG.

Grouping promoters into two separate classes largely eliminated the high predictive ability of CG dinucleotide on ubiquitous promoter activity: for HCG promoters, $R^2=0.048$ and AUC = 60.8%; for LCG promoters, $R^2=0.25$ and AUC=77.5%. In all subsequent analysis, I consider the performance of CG as the baseline against which the motif predictions are compared.

### 3.7 Predicting ubiquitous and cell line-specific active promoters

Using support vector machines (SVM), I designed discriminative models to predict whether a promoter would be active or inactive given a set of TF motifs in that promoter. Each of the 691 TF motifs considered were represented by a composite score computed by the algorithm Clover (Frith et al., 2004a). Using the binarized promoter activity scores as previously described, I designated promoters that were active in all eight cell lines as ubiquitous, and promoters that were active in only one cell line as tissue-specific. I assume that each promoter set is regulated by a specific set of TF motifs, and implement separate models for each promoter set. The number of promoters in each model is included in column 3 of Table 3-1. Promoter sets that had fewer than 20 promoters were not considered because there would be insufficient instances to produce a robust model evaluation criterion. In total, I considered 9 models: 2 models for HCG promoters (Ubiquitous and HeLa-specific)
Figure 3.7: Distribution of CG dinucleotide content in 4575 promoters. Normalized CG is a ratio of observed and expected number of CG dinucleotides. Low-CG promoters have normalized CG < 0.5, whereas High-CG promoters have normalized CG ≥0.5.
Table 3.1: AUCs of best performing models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of promoters</th>
<th>% AUC of best model</th>
<th>% AUC of best model − %AUC (CG)</th>
<th>Number of motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCG HeLa</td>
<td>43</td>
<td>99.7</td>
<td>45.6</td>
<td>112</td>
</tr>
<tr>
<td>HCG Ubiquitous</td>
<td>201</td>
<td>93.1</td>
<td>32.3</td>
<td>27</td>
</tr>
<tr>
<td>LCG U87mg</td>
<td>77</td>
<td>89.6</td>
<td>41.1</td>
<td>163</td>
</tr>
<tr>
<td>LCG Ags</td>
<td>61</td>
<td>79.1</td>
<td>11.7</td>
<td>37</td>
</tr>
<tr>
<td>LCG HepG2</td>
<td>29</td>
<td>98.6</td>
<td>57.8</td>
<td>135</td>
</tr>
<tr>
<td>LCG HeLa</td>
<td>109</td>
<td>82.9</td>
<td>28.5</td>
<td>23</td>
</tr>
<tr>
<td>LCG Hct116</td>
<td>44</td>
<td>87.4</td>
<td>31.8</td>
<td>81</td>
</tr>
<tr>
<td>LCG T98g</td>
<td>33</td>
<td>95.7</td>
<td>53.9</td>
<td>50</td>
</tr>
<tr>
<td>LCG Ubiquitous</td>
<td>70</td>
<td>92.5</td>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>

The model evaluation criterion I used was the AUC, which balances sensitivity and specificity. To avoid over-fitting, I performed 5 cross-validation trials and the average AUC for the best performing model in each category is reported in Table 3-1. Because the combinatorial space of all examined TFs would be intractable to traverse, I first ranked the TFs by their individual AUCs, and then added them sequentially to form a cumulative model. I attempted more complex feature selection strategies such as Recursive Feature Elimination (Guyon et al., 2002), and RSVM (Zhang et al., 2006), but they produced overall lower ROC scores than those produced by simply adding motifs according to their ranks (Fig 3.8). Overall, all 9 models performed well, with an average AUC of 92.8% for the ubiquitous models, and 90.4% for the cell line-specific models.
Figure 3-8: Comparison of TF feature selection by RSVM and univariate ranking. TF combinations were selected by Recursive Feature Selection for Support Vector Machines (RSVM), but did not produce models with higher AUC than models where TF combinations were selected by sequentially adding TFs according to their performance in univariate tests.
Interestingly, ubiquitous models required fewer TF motifs to achieve a high AUC than cell-line specific models. Only 27 motifs were required in the HCG ubiquitous model, and only 6 motifs were required in the LCG ubiquitous model. In contrast, the number of motifs required in the cell-line specific models ranged from 23 to 163 motifs, with an average of 89 motifs per model (column 6 in Table 3.1). This suggests that the complex regulation of cell line-specific promoter activity requires many more TFs, whereas a few key TFs can adequately regulate ubiquitous promoters.

3.8 Cis-regulatory module discovery

Identifying TFs that most influence promoter activity is an important step in understanding transcription regulation. I found that the models maintained high predictive ability even after narrowing down to modules of four motifs. The performance of these modules exceeded that of CG dinucleotide counts by an average of 23.2% AUC (column 5 Table 3-2).

Each model identifies unique subsets of TFs that regulate cell line-specific promoter activity (Table 3.2). Several TFs that are known to regulate activity in a specific tissue type were recapitulated in the appropriate models. For example HNF4α in liver (Watt et al. 2003) and CREB in brain (Gass and Riva, 2007; Han et al., 2007; Mantamadiotis et al., 2002) were specific to the HepG2-specific model, and the T98g model respectively (Table 3.2). Half (13/26) of the motifs in the 4-motif modules corresponded to TFs that garnered literature support for tissue specificity, and 19.1% (5/26) of the motifs were potentially novel as they did not associate with a known TF.

3.9 Predicting ubiquitous and cell line-specific expressed genes

Computational strategies to discover TFs that regulate tissue-specific transcription have typically relied on gene expression data (Davies et al., 2007; Smith et al., 2006; Smith et al., 2007). As a comparison to the previous results studying promoter activities I also conducted parallel analyses using genome-wide gene expression data. Overall, the models
<table>
<thead>
<tr>
<th>Model</th>
<th>% AUC (module)</th>
<th>% AUC (module) - %AUC (CG)</th>
<th>ETS-family</th>
<th>Nrf-1</th>
<th>Stat1</th>
<th>Spl1</th>
<th>ACTWSNACTNY</th>
<th>E2F</th>
<th>EGR</th>
<th>FXR-RXR</th>
<th>Pax4</th>
<th>FOX</th>
<th>MAF</th>
<th>OCT1</th>
<th>Pax2</th>
<th>RAR</th>
<th>TNCATNTCCYR</th>
<th>TGCTGAY</th>
<th>TGACATY</th>
<th>CREB</th>
<th>RSCRE1</th>
<th>MEF2</th>
<th>CACCC-BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCG HeLa</td>
<td>90</td>
<td>36.8</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCG Ubiquitous</td>
<td>85</td>
<td>24</td>
<td>X*</td>
<td>X*</td>
<td>X*</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCG U87mg</td>
<td>71</td>
<td>22.5</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCG Ags</td>
<td>74.3</td>
<td>6.9</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCG HepG2</td>
<td>75.7</td>
<td>34.9</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCG HeLa</td>
<td>74.3</td>
<td>19.9</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCG Hct116</td>
<td>68</td>
<td>12.4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCG T98g</td>
<td>79.5</td>
<td>37.7</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCG Ubiquitous</td>
<td>90.9</td>
<td>13.4</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X*</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Motif modules discovered using the promoter activity dataset. Asterisk (*) indicates literature support.
trained on promoter activity data performed better than models trained on gene expression data. The average AUC above CG dinucleotide content was 35.3% in the promoter activity models, compared to 26.1% in the gene expression models. This is most pronounced in the ubiquitous models where the performances of the TF modules above and beyond CG dinucleotide content were 24% and 13.4% for the ubiquitous HCG and LCG models respectively (Table 3.2). In contrast, the performances of the same models trained on gene expression data did not exceed 6% (Table 3.3).

Comparing the literature support of specific TF motifs, I found that only 28% (9/32) of the TFs identified in the gene expression analysis were supported by literature, compared to 50% (13/26) in the promoter activity analysis. It is likely that the promoter activity assay is more appropriate for sequence-based modeling than gene expression because the promoter activity assay directly measures the activity of a known fragment of DNA, whereas endogenous expression can be additionally confounded by chromatin or other epigenetic mechanisms.

3.10 Methods

3.10.1 Promoter transient transfection reporter assay

Promoter transient transfection reporter assays were performed as described previously, with some modifications. All promoter constructs were tested in triplicate across each of the eight cell lines. Each plate also contained four known positive controls across a range of activities, as well as four known negative controls, also across a range of activities. The 5' reporter constructs were custom prepared by SwitchGear Genomics (Menlo Park, CA). The fragments are ”promoter-plus” 5’ sequences that are, on average, 1kb in size. Sequences are PCR amplified from human genomic DNA and then subcloned into the pGL4 reporter vector (Promega). Each 5' reporter construct is sequence-validated.

The luciferase signal was divided by the mean of the negative controls in a given cell type and plate. This controls for the plate-to-plate variations in transfection efficiency and also normalizes the promoter activity such that it is a measure of the fold increase in
<table>
<thead>
<tr>
<th>Model</th>
<th>% AUC (module)</th>
<th>% AUC (module) - % AUC (CG)</th>
<th>EFS-family</th>
<th>E2F</th>
<th>CCGGNN/NACC</th>
<th>Whn</th>
<th>XGCG/GYBGG</th>
<th>NE/YN</th>
<th>NRF-1</th>
<th>MAZI</th>
<th>RERB</th>
<th>RACTN/RXR</th>
<th>TNC</th>
<th>RERE</th>
<th>CRE/BRF</th>
<th>PAX4</th>
<th>ACCTT/CTC</th>
<th>Tax/CREB</th>
<th>Fosl</th>
<th>ADAR</th>
<th>MIR</th>
<th>TSA</th>
<th>ZNF1</th>
<th>GATA</th>
<th>PPR</th>
<th>ZCE5</th>
<th>CCRG</th>
<th>GGGCAR</th>
<th>GGGCAR</th>
<th>CCGN/AGRKGC</th>
<th>AHR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCG U87mg</td>
<td>77</td>
<td>2.6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82.6</td>
<td>X</td>
</tr>
<tr>
<td>HCG Ags</td>
<td>79.5</td>
<td>5.6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80.3</td>
<td>X</td>
</tr>
<tr>
<td>HCG HepG2</td>
<td>81.6</td>
<td>3.7</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>75.7</td>
<td>X</td>
</tr>
<tr>
<td>HCG HeLa</td>
<td>87.7</td>
<td>25.6</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>HCG Hct116</td>
<td>72.4</td>
<td>4.9</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>HCG T98G</td>
<td>78.1</td>
<td>39.3</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>HCG G402</td>
<td>79</td>
<td>12</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>HCG Ubiquitous</td>
<td>84.6</td>
<td>-0.2</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG U87mg</td>
<td>80.1</td>
<td>34.5</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG Ags</td>
<td>85.3</td>
<td>18.8</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG HepG2</td>
<td>75.7</td>
<td>36.4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG HeLa</td>
<td>73</td>
<td>23</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG Hct116</td>
<td>75.4</td>
<td>12.3</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG T98G</td>
<td>77.6</td>
<td>3.3</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG G402</td>
<td>80.8</td>
<td>17</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>LCG Ubiquitous</td>
<td>82.6</td>
<td>-1</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Motif modules discovered using the gene expression dataset. asterisk (*) indicates literature support.
activity over background. A threshold was imposed at the trough of the distribution of log2 promoter activity scores, corresponding to a value of 0. Positive promoters therefore have a luciferase signal that exceeds background.

3.10.2 Whole genome expression analysis

Total RNA was prepared with Trizol (Invitrogen) according to the manufacturer’s protocols. Three biological replicates were prepared for each of the eight immortal cell lines used in this study. Purity of the RNA was assessed by NanoDrop. RNA was amplified using standard procedures from the Illumina TotalPrep RNA amplification kit. Illumina HumanRef−8 v2 Expression BeadChips were used for whole genome expression analysis on each of the three biological replicates. Gene identifiers from all of the transcripts on the BeadArrays were matched to promoter predictions from SwitchGear Genomics (www.switchdb.com; score > 20) to find transcripts that were not associated with alternative promoters.

3.10.3 CG dinucleotide count normalization

I first counted the number of CG and C+G counts in a promoter and then divided each by the length of the promoter. The normalized CG dinucleotide count was computed as $\frac{CG_{\text{observed}}}{CG_{\text{expected}}}$ defined in Saxonov et al. Expected CG dinucleotide was defined as $((G + C)/2)^2$.

3.10.4 SVM implementation details

For each promoter class (HCG or LCG), the most abundant profile was chosen as the background promoter set and the foreground promoter sets were either ubiquitous promoters, or cell line specific promoters that were active in only one cell line. For example, one model was trained to discriminate between ubiquitously active LCG promoters (foreground) and ubiquitously inactive LCG promoters (background). Another model was trained to discriminate between HeLa-specific LCG promoters and ubiquitously inactive LCG promoters (the same background promoter set). Some of the foreground promoter sets contained less than
20 promoters, which was not enough to produce robust AUC scores averaged across 5-fold cross validated trials; 80 percent of the data was used to train the SVM model, and 20 percent of the data was reserved for testing. These cases could not be reliably tested and were discarded, leaving 2 models in the HCG promoter class (Ubiquitous and HeLa-specific) and 7 models in the LCG promoter class (Ubiquitous, U87mg-, Ags-, HepG2-, HeLa-, Hct116-, T98G-, and G402-specific).

All models were implemented in R 2.5.0, using the e1071 package interface to the libsvm algorithm written by Chih-Chung Chang and Chih-Jen Lin. A linear kernel was used which did not require any arbitrary parameters.

3.10.5 Reducing redundancy in motif sets

I used the program Malign (Haverty et al., 2004), to prune out redundant motifs that had pairwise Pearson correlations greater than 0.2. I then pick the motif with the best rank in each model to represent the group of redundant motifs.
Chapter 4

Single-base-resolution Map of TF Footprints in Human Promoters

4.1 Introduction

Transcription factors (TFs) play a crucial role in all kinds of biological processes by dynamically inducing gene expression in response to environmental stimuli and developmental conditions. Numerous studies have shown that TFs are required for correct cell fate specification, regulating the production of tissue specific genes throughout development (Gangenahalli et al., 2005; Oakes et al., 2008; Singh et al., 2005; Strumpf et al., 2005). Recognizing the global importance of TF regulation, genome-wide maps of TF binding locations are being produced for a handful of TFs in various conditions and cell-types (Bieda et al., 2006; Johnson et al., 2007; Robertson et al., 2007). However, in general, studies of TF binding reflect a tradeoff between throughput and resolution. High-throughput Chromatin immuno-precipitation studies produce target lists about 1kb in size, while traditional mobility shift (Fried and Crothers, 1981; Garner and Revzin, 1981) or DNA footprinting (Dynan and Tjian, 1983; Galas and Schmitz, 1978) assays identify the exact 6-10 base TF footprint on a site-by-site basis. Despite these caveats, these methods have made significant strides toward the understanding of TF dependent gene expression.

Databases such as Transfac (Matys et al., 2006) and Jaspar (Vlieghe et al., 2006) compile TF binding footprints from multiple experiments to construct motifs or position-specific scoring matrices (PSSM), representing the binding affinities of TFs according to a thermodynamic model (Stormo, 2000). PSSMs can then be used as seeds in various computational algorithms to predict TF footprints in uncharacterized regions. Some of these
algorithms further rely on information such as conservation (Xie et al., 2005), or overlap with ChIP target lists (Holloway et al., 2005), although the marginal contribution by each additional information source is unknown. Regardless, any computational algorithm would suffer if it is trained on poor quality PSSMs, or validated by an incomplete collection of TF footprints. Therefore, a high-throughput, high-resolution map of TF footprints would be an important step in characterizing TF function in general, supporting the development of more precise models of TF binding.

In order to create a single-base resolution map of TF binding footprints, motifs that could predict active and tissue-specific promoter activities across eight cell lines were collected from a previous study (Johnson et al., 2008 in preparation). Support vector machine models were constructed to differentiate between ubiquitous and tissue-specific promoter activity using quantitative scores representing the presence of 691 TF motifs in approximately 5000 promoters. Twenty-one most predictive TF motifs were prioritized in the current study. Eighty-four high-quality TF footprints were selected for preliminary testing. A simple sequence-based strategy was employed to select the strongest TF footprint on a given promoter. Filtering parameters were fine-tuned after each round of testing to increase the chances of identifying functional TF binding footprints. Additional filtering criteria based on conservation, chromatin state, and overlap with ChIP target lists were also introduced to evaluate the relative importance of DNA sequence, conservation, and chromatin effects on TF binding. This work critically evaluated biological signatures that inform TF binding in Human promoters, and serves as a foundation for mapping functional TF footprints genome-wide.

4.2 TF binding footprints in Human promoters

The binding potential of TFs are first characterized at every position and orientation along a promoter using the program Possum. Possum implements the standard method of scoring by PSSMs (Stormo, 2000; Wasserman and Sandelin, 2004) and reports the log likelihood ratio of observing a binding site given the nucleotide frequencies in the chosen PSSM, versus
the probability of observing a binding site given the nucleotide frequencies in a background model. Putative TF binding footprints are then screened via three filters described in the following sections.

4.3 Filtering criterion 1: TF binding site must be solitary

Transcription factor binding sites can appear many times along a promoter, such that secondary TF binding sites may compensate for a primary TF binding site if the primary site is damaged or mutated. As I proposed to mutate one TF binding site at a time, I wanted to reduce the chances of having a TF bind to a secondary site after mutating the primary site. The first filtering criterion requires that the difference between the site with the highest Possum score and the site with the second highest Possum score be in the top 20% of a background distribution. The background distribution was calculated by surveying Possum score difference among 246 promoters in the ENCODE region. Results from these single-site experiments would be instrumental in the future development of higher-order mutagenesis experiments, where two or more sites will be mutated to tease out the individual and combinatorial effects of multiple TF binding sites.

4.4 Filtering criterion 2: The highest scoring TF binding site must be statistically significant

To generate a background distribution, Possum scores for every position along a training set of 4575 Human promoters were computed for 21 selected TFs. The example in Figure 4.1 shows the distribution of GABPA Possum scores. The filtering criterion requires that GABPA binding sites selected for mutagenesis have Possum scores beyond the 99.95th percentile in the distributions. This corresponds to a Possum score of 4.84 for low CpG, and a Possum score of 6.36 for high CpG promoters (vertical lines in Fig. 4.1). Depending on the results of the first batch of mutagenesis experiments, I will tighten or loosen the p-value cutoff for the next round of TF binding site picks.
Figure 4.1: Distribution of GABPA Possum scores in Human promoters. Possum Score cutoff is different for High CG (HCG) and Low CpG (LCG) promoters. The cutoff was set to the 99.95% quantile.
4.5 Preliminary mutagenesis of five TF footprints

As a preliminary test, I selected five GABPA binding sites to test by mutagenesis and transient transfection assay. In previous studies, I had discovered that GABPA was an important regulator of ubiquitous promoter activity in both HCG and LCG promoters, and was also implicated in the regulation of bidirectional promoters. From these studies, I had a high-quality PSSM for GABP, and knew that it was highly predictive of \textit{in vivo} binding. Patrick Collins at SwitchGear Genomics mutated the predicted GABPA footprints by random and transversion mutations of the wild-type bases. Results in Figure 4.2 indicate all five predicted sites are functional.

4.6 Possible filtering criterion: Conservation

As the conservation of DNA sequence through evolutionary time is a strong indicator of functional TF binding sites and other important regulatory regions in the genome, one possible filtering criterion to consider is the degree of alignment among species from different branches on the evolutionary tree. From the start, I did not incorporate conservation information in the TF binding site predictions because this information is not available in the cell, and TFs do not use conservation to guide binding preferences \textit{per se}. However, many groups have been successful in exploiting multiple-sequence alignment information to guide TF binding site predictions (Xie et al., 2005), as constrained regions tend to co-occur with open chromatin marks, CpG islands, and other features may be favorable for TF localization. In Figures 4.3 to 4.7, I examine five GABPA binding footprints tested previously and find that all five functional footprints are not conserved beyond mammals. Conservation patterns in these instances suggest that multiple-sequence alignments among mammals would be sufficient for predicting TF function.
Figure 4-2: Mutagenesis of five GABPA binding footprints. The promoter activity of five wildtype promoters was first assayed by transient transfection, and assayed again after random (blue) and transversion (purple) mutation of a GABPA binding site in each of the promoters. The change in promoter activity is plotted in log scale, showing that all five GABPA footprints resulted in diminished promoter activity. All five predicted GABPA binding sites were functionally active.
Figure 4.3: Conservation of a GABPA binding footprint (CHR2_P1199_R1). The motif is conserved up to Dog.

<table>
<thead>
<tr>
<th>chr2</th>
<th>22456755</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>C</th>
<th>220261760</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaps</td>
<td>human</td>
<td>mouse</td>
<td>rat</td>
<td>dog</td>
<td>armadillo</td>
<td>elephant</td>
<td>opossum</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
</tbody>
</table>

Simple Nucleotide Polymorphisms (dbSNP build 125)
Repeating Elements by RepeatMasker

Figure 4.4: Conservation of a GABPA binding footprint (CHR7_M0172_R1). The motif is conserved up to Rat.

<table>
<thead>
<tr>
<th>chr7</th>
<th>26643858</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>26493865</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaps</td>
<td>human</td>
<td>mouse</td>
<td>rat</td>
<td>rabbit</td>
<td>dog</td>
<td>armadillo</td>
<td>elephant</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
</tbody>
</table>

Simple Nucleotide Polymorphisms (dbSNP build 125)
Repeating Elements by RepeatMasker
Figure 4.5: Conservation of a GABPA binding footprint (CHR21_M0109_R1). The motif is conserved up to Rat.

Figure 4.6: Conservation of a GABPA binding footprint (CHR22_P0237_R3). The TTCC consensus is conserved up to Armadillo.
4.7 Possible filtering criterion: Promoter overlap with ChIP target lists

Chromatin immuno-precipitation (ChIP) combined with tiling arrays (ChIP-chip) or sequencing (ChIP-seq) is a powerful *in vivo* method to identify protein-bound DNA fragments. To cross-check the TF predictions, I overlapped ChIP-chip datasets of relevant TFs with promoter subsets that were predicted by models constructed in Chapter 3. In Figure 4-8, the fraction of 246 ENCODE promoters that overlap with each ChIP-ed factor is plotted in the left-most blue bar as a baseline. By comparison, ubiquitously active HCG promoters (second-to-left red bar) are enriched for cMyc, E2F1, E2F4, and Sp1 binding, corresponding to motif ranks of 31, 4, 48, and 3 respectively from the ubiquitously active HCG model presented in Chapter 3. Factors that were ChIP-ed tended to be general TFs that function in many cell-types and conditions. As the technology gains resolution and expands the TFs studied, biologists can begin to use ChIP datasets to filter the TF binding footprint predictions.
Figure 4.8: Overlap of Human promoters with ChIP target lists.
Chapter 5

Conclusion

5.1 Summary

In this dissertation I explored DNA sequence and epigenetic features that drive Human promoter function. Major regulators such as TFs, modified histones, DNA methylases and demethylases have been known for decades, but much of their regulatory complexity remains poorly understood. Although components in the transcriptional machinery can potentially interact in an astronomically large combinatorial space, scientists are methodically uncovering the regulatory mechanisms governing gene expression in a number of temporal and spatial conditions. In particular, I presented detailed analyses of regulatory elements found in bidirectional promoters and promoters that have tissue-specific activity. The transcription factor GABPA was an important regulator in both studies, highlighting the possibility of reusing and repurposing transcriptional components for different functions. I discovered that the recognition sequences for GABPA, MYC, E2F1, E2F4, NRF-1, NFY, YY1, and ACTACAnnTCC were overrepresented in bidirectional promoters, while the vast majority (73%) of vertebrate transcription factor motifs are underrepresented. Similarly, tissue-specific transcription regulation is largely determined by a few key TFs. Computational models were trained to rely on only 4 TFs (compared to the full set of 691 known TFs) without sacrificing much predictive power. Transcription regulatory rules can indeed be simplified, as laws of nature have proven to be elegantly simple. This gives us hope that we may eventually understand the world around us through systematic study.
5.2 Future directions

5.2.1 Integrated models of transcription regulation

By careful selection of experimental and computational analysis methods, I deconvoluted the DNA sequence and epigenetic regulatory information encoded in a large set of human promoters. Significantly, I matched the appropriate predictors (sequence features) with the appropriate predictand (promoter activity). The transient transfection assays measured the promoter activity of a known fragment of DNA without epigenetic modification, a feature that matched our sequence-based modeling strategy. As more information is generated in the field, we can begin to look more broadly and build more comprehensive models of gene regulation, incorporating chromatin, long range elements, and epigenetic features. These types of integrated models have been implemented previously, matching predictors such as k-mer counts, ChIP of TFs, and ChIP of modified histones, to gene expression, which is the most appropriate predictand (Holloway et al., 2005). I showed that sequence features explain approximately 23%, while DNA methylation explains an additional 5% of the variation observed in endogenous gene expression. This analysis demonstrates the importance of computational modeling in revealing the quantitative contributions of independent biological features, and would not have been possible if high-throughput datasets were not systematically generated to allow analyses in concordant tissue types and environmental contexts. This strategy was adopted by the ENCODE project (Birney et al., 2007), setting the stage for a new era in genomics (Weng and Guigo, 2008), where disparate components that function together can be measured separately but analyzed in conjunction to elucidate novel biology.

5.2.2 Synthetic promoters

The findings presented in this dissertation can easily be applied to the design of synthetic promoters. In a subsequent study, Collins et al. was able to generate bidirectional promoter activity from unidirectional promoters after artificially introducing GABPA binding sites in the unidirectional promoters (Collins et al., 2007). Most synthetic promoters are
designed by randomly mutating bases followed by massive screening to produce a handful of promoters with desired function (Kagiya et al., 2005). Understanding the sequence elements that drive human promoter function opens up the possibility for rational design of synthetic promoters. Synthetic bidirectional promoters are already being investigated for their ability to introduce multiple transgenes for gene therapy (Amendola et al., 2005). Synthetic promoters are also used to optimize metabolic pathways for the production of biofuels (Jensen and Hammer, 1998). Additional applications of synthetic promoters will likely emerge as we gain a deeper understanding of the DNA sequence and epigenetic elements that drive promoter function.
References


CURRICULUM VITAE

Jane Marie Landolin

janemlin@bu.edu
(617) 763-9022
50 Johnstone Drive Apt 206
San Francisco, CA, 941315

EDUCATION

Boston University, Boston, MA
Doctor of philosophy in Biomedical Engineering, 2009
Dissertation: DNA Sequence and Epigenetic Features that Drive Human Promoter Function

University of California - Los Angeles, Los Angeles, CA
Master of Science in Biomedical Engineering, 2003
Thesis: Model discrimination analysis of GLUT4 trafficking and intracellular sorting
Bachelor of Science in Cybernetics, specializing in bioinformatics, 2003

RELEVANT COURSEWORK

Biology: Genetics, Molecular Bioengineering (gel electrophoresis, PCR, EMSA)
Computer: Advanced Programming (C++), Data Structures and Algorithms, Database Systems

EXPERIENCE

Boston University
Department of Biomedical Engineering 2005 - present
Research Associate

• Identifying differentially expressed genes from microarray gene expression experiments using R

• Designing gene regulation models using packages such as Bioconductor, limma, and e1070 in R
• Analyzing and aligning sequence using Perl, and bioinformatics tools such as BLAST, BLAT and ClustalW

• Discovering transcription factor binding sites using tools such as MEME, Clover, and YMF

• Identifying functional elements from chIP-chip, chIP-seq, and luciferase transfection experiments

• Summarizing and visualizing multi-dimensional data using R

• Access, query, and data mine bioinformatics databases such as Genbank, UCSC Genome Browser, TRANSFAC, SNPdb, GEO, and PubMed

University of California - Los Angeles
Department of Computer Science 2002 - 2003
Research Assistant

• Designed and simulated kinetic models of insulin signaling using VISSIM for diabetes research

• Designed and simulated compartmental models of glucose transporter translocation using SAAMII

University of California - Los Angeles
Genotyping and Sequencing Core 2001 - 2002
Database Program Analyst

• Designed and implemented custom entry forms for an in-house Microsoft Access relational database

• Designed and implemented database query tools to streamline data acquisition-to-analysis pipeline using Visual Basic

SKILLS

Computer: Linux, Macintosh, Windows, Portable Batch System for parallel computing
Programming: R, C++, Perl, BioPerl, Wiki, XML, HTML, CSS, SQL
Engineering: Matlab, SAS, VISSIM, SAAMII
Languages: Mandarin, Cantonese

PUBLICATIONS


* Lin JM is an author in the transcription regulation group


Johnson DJ*, Lin JM*, Weng Z, Myers RM. Sequence features that drive human promoter function and tissue specificity. (In preparation for Genome Research) *co-first authors

Lin JM, Trinklein ND, Collins PC, Myers RM, Weng Z. Functional annotation of transcription factor binding footprints at base-pair resolution. (In preparation)

AWARDS
Research Fellow in Quantitative Biology and Physiology (NIH) 2003-2005
Jo-Belle Wolf Undergraduate Scholarship 2002-2003
Irving and Jean Stone Undergraduate Research Award 2002
IGERT Undergraduate Bioinformatics Summer Internship (NSF)
Project: Metabolic and proteomic evaluation of cellular signaling 2002
Project: High throughput genomic data analysis 2001
Departmental Scholar (combined Bachelor and Master degrees) 1998-2003

TEACHING
EK424: Thermodynamics and Statistical Mechanics Spring 2004

EXTRACURRICULAR AND COMMUNITY
Boston University graduate student mentor 2004-2008
Facilitator in Responsible Conduct of Research Training Program 2006-2007
Senator for Student Association of Graduate Engineers (SAGE) 2004-2005
Cybernetics student mentor 2000-2003