UNIVERSITY OF CALIFORNIA, SAN DIEGO

IDENTIFICATION AND CHARACTERIZATION OF TRANSCRIPTIONAL ENHANCERS IN THE HUMAN GENOME

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Biomedical Sciences

by

Nathaniel David Heintzman

Committee in charge:

Professor Bing Ren, Chair
Professor Webster Cavenee
Professor Frank Furnari
Professor Chris Glass
Professor Jean Wang

2007

The dissertation of Nathaniel David Heintzman is approved, and is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2007

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Many thanks to my advisor and mentor, Dr. Bing Ren, for his extensive contributions to my incredible graduate experience. Thanks to many members of the Ren lab, especially Rhona Stuart and Lindsey Harp for their indispensable help in conducting my experiments after my shoulder injury; and Gary Hon, Dr. Leah Barrera, and Dr. Keith Ching, without whose computational expertise my project would have languished.

Thanks to certain members of neighboring labs, past and present, for helping me find levity in the midst of long days and nights of experiments and analysis, especially Dr. Tim Fenton and Sharsti Sandall.

Thanks to Dr. Greg Crawford for inventing DNase-chip at the perfect time and for being so generous with his data and resources. Thanks to Roland Green and the whole crew at NimbleGen Systems, Inc., for accelerating my work along the microarray superhighway. Thanks to Dr. Webster Cavenee for his candid advice and exemplary leadership, to the faculty and staff of the Ludwig Institute for Cancer Research for creating an environment where I could focus on doing great science, and to my committee members for their invaluable advice along the way.

Also essential to my success as a graduate student are my exceptionally supportive family: Mom, Dad, Sarah, Matthew, and Rebecca; my great friends in San Diego and across the nation, especially David Holmberg, Jr. and Peter H. Nerothin; and my colleagues in Biomedical Sciences and other graduate programs.

Thanks to Gina Butcher and Leanne Nordeman, the BMS bosses, for taking care of business, and to the various chairs of BMS through the ages for their incredible effort and success in running an elite graduate program.

Thanks to Porter's Pub for facilitating out-of-the-box thinking when I needed it most. Thanks to the government and citizens of Mexico for making it so easy to camp and surf in Baja. Thanks to Dr. Eric Skoblar and Dr. Robert Pedowitz for patching me up so I can once again enjoy such pursuits.

Thanks to my undergraduate professors Dr. Ellis Bell, Dr. John Lammert, and Dr. Colleen Jacks for guiding me toward graduate school, and to Dr. J. Craig Venter for guiding me to UCSD in particular (and also for the human genome sequence, without which my project would have been utterly impossible; along those lines, thanks also to Dr. Francis S. Collins).

Thanks to Dr. Will Alaynick, Omai Garner, Dr. Ravi Kolla, and our adoring fans for making Mystykk Stallion such a fulfilling distraction.

Chapter 1, in full, is a reprint of the material as it appears in "The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome" in Cellular and Molecular Life Sciences 2007, Heintzman ND and Ren B. The dissertation author was the primary investigator and author of this publication.

Chapter 2, in full, is a reprint of the material as it appears in "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nature Genetics 2007, Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, and Ren B. The dissertation author was the primary investigator and author of this publication.

Chapter 3, in full, is a reprint of the material as it appears in a manuscript submitted to Nature in June 2007, “A Genome-wide Map of Human Transcriptional Enhancers.” Heintzman ND, Hon GC, Kheradpour P, Stark A, Ching KA, Stuart RK, Harp LF, Hawkins RD, Ching CW, Liu H, Zhang X, Green RD, Crawford GE, Kellis M, and Ren B. The dissertation author was the primary investigator and author of this publication.

VITA

2001 B.A. Biochemistry and Molecular Biology
Gustavus Adolphus College, St. Peter, MN

2007 Ph.D. Biomedical Sciences
University of California San Diego, La Jolla, CA

PUBLICATIONS

Heintzman ND, Ren B. “The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome.” Cellular and Molecular Life Sciences, Vol.64(4):386-400, 2007.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, and Ren B. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.” Nature Genetics, Vol.39(3):311-8, 2007.

ENCODE project consortium. “Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.” Nature, Vol.447(7146):799-816, 2007.

Heintzman ND, Hon GC, Kheradpour P, Stark A, Ching KA, Stuart RK, Harp LF, Hawkins RD, Ching CW, Liu H, Zhang X, Green RD, Crawford GE, Kellis M, and Ren B. “A genome-wide map of human transcriptional enhancers.” Submitted.

ABSTRACT OF THE DISSERTATION

IDENTIFICATION AND CHARACTERIZATION OF TRANSCRIPTIONAL ENHANCERS IN THE HUMAN GENOME

by

Nathaniel David Heintzman

Doctor of Philosophy in Biomedical Sciences
University of California, San Diego 2007

Professor Bing Ren, Chair

This dissertation describes the use of high-throughput molecular biological techniques and bioinformatics to systematically locate and characterize transcriptional enhancers in the human genome. Enhancers are an important class of transcriptional regulatory elements, along with promoters, silencers, insulators, and locus control regions. Though critical to proper regulation of gene expression in space and time, enhancers have been difficult to locate in the human genome due to their widespread distribution and poorly understood sequence features. In this work, I discuss the discovery of physical features of enhancers that allow their distinct identification throughout the genome of human cells and the insights gained by the first genome-wide analysis of human transcriptional enhancers.

Chapter 1 introduces principles of transcriptional regulation and summarizes high-throughput technologies employed to locate transcriptional promoters.

Chapter 2 details the discovery of distinct chromatin signatures for promoters and enhancers and the development of novel computational strategies to predict these regulatory elements in 1% of the human genome.

Chapter 3 describes the extension of the enhancer prediction model to the entire human genome and insights gained from large-scale analysis of the characteristics of these enhancers.

Chapter 4 discusses the development of a related high-throughput method, RiGS, designed to facilitate functional genomic screens in mammalian systems.

Chapter 1

The Gateway to Transcription: Identifying, Characterizing, and Understanding Promoters in the Eukaryotic Genome

Abstract

Eukaryotic transcriptional regulation requires the integration of complex signals by the transcriptional promoter. Distinct sequence elements, characteristic chromatin modifications and coordinated protein-DNA interactions at these sequences constitute a transcriptional regulatory code that remains poorly understood today. Here, we review recent experimental and computational advances that have enabled the identification and analysis of transcriptional promoters on an unprecedented scale, laying a foundation for systematic determination of the transcriptional regulatory networks in eukaryotic cells. The knowledge gained from these large-scale investigations has challenged some conventional concepts of promoter structure and function, and provided valuable insights into the complex gene regulatory mechanisms in a variety of organisms.

1.1 Introduction

Regulation of gene expression in eukaryotes requires precise spatial and temporal coordination of a multitude of general and specific transcription factors at *cis*-regulatory elements, including enhancers, silencers, insulators, and promoters[1-3]. Recognition and binding of these sequences by transcription factors occurs within the context of chromatin, whose dynamic structural characteristics play a significant role in regulating gene expression[4]. The histone proteins that underpin chromatin structure are subject to an ever-expanding variety of covalent modifications that serve as the result of signaling pathways, as epigenetic markers for cellular events, and as molecular beacons for additional modifying enzymes and transcriptional regulators that influence chromatin architecture and gene expression[5]. The transcriptional promoter is the nexus of all of these levels of regulation, serving as the ultimate determinant in the transcription of any gene by integrating the manifold influences of DNA sequence, transcription factor binding, epigenetic features, and signal transduction events. Understanding the mechanisms by which promoters integrate these regulatory inputs is critical to our comprehension of transcriptional regulation in human evolution, development, disease, and environmental response.

Eukaryotic promoter structure and regulation of expression for protein-coding genes have been extensively reviewed elsewhere[1,2,6], so we will briefly define and summarize key features and events involved in regulating the initiation of transcription (see Figure 1.1). A eukaryotic promoter is located at the 5’ end of its transcribed sequence and serves as the point of transcriptional initiation. Typically, the term "promoter" refers to the "core promoter" and its adjacent sequences. The core promoter immediately surrounds the transcription start site (TSS) and comprises 70-80 base pairs that contain canonical sequence features (described below) sufficient for recognition by the basal transcriptional machinery and initiation of transcription. The "proximal promoter" includes the region extending upstream of the core promoter (generally ~250 bp from the TSS[7], though this limit can be somewhat subjective). Proximal promoters contain other sequence features critical to transcriptional regulation, for instance binding sites for tissue-specific transcription factors, and may in fact encompass transcriptional enhancers (which by their nature impart additional regulatory specificity to expression of the target gene), but due to their close proximity to the core promoter and our evolving understanding of promoter structure we will refer to these regions collectively as the promoter unless otherwise noted.

To prepare a promoter for transcriptional initiation, sequence-specific transcription factors bind to regulatory sites in the promoter and enhancers, recruiting coactivators such as chromatin remodeling enzymes and histone modifying enzymes that alter nucleosome structure and position. Diverse protein complexes are involved in this process[8,9]. The precise timing and ordering of these events is still debated, but the end result is a regulated reorganization of chromatin structure within the promoter. This restructuring permits and stabilizes binding of the basal transcriptional machinery, composed of RNA polymerase II (RNAPII) and numerous general transcription factors required for proper positioning of the polymerase and interactions with other specific regulatory proteins. Poised to begin transcription, this structure is referred to as the Pre-Initiation Complex (PIC). The PIC interacts with a variety of additional regulatory proteins, such as the Mediator complex[10], involved in structural and temporal regulation of initiation. Through a poorly understood mechanism, the 11-15 bp of DNA around the TSS "melts" to allow positioning of the template strand within the active site of RNAPII, and transcription begins. After ~30 nt of RNA have been transcribed, RNAPII physically separates from the promoter and the rest of the PIC and enters the transcriptional elongation phase, now associating with different regulatory factors that influence processive and accurate RNA synthesis and chromatin remodeling. The precise mechanisms of these events are still being actively researched. For example, recent evidence suggests that transient double-strand breaks in the DNA at promoters is required for regulated transcription[11], and other studies have begun to dissect the epigenetic events responsible for selective chromatin opening at active promoters, distinct from the chromatin remodeling that occurs in the coding region during elongation[12].

Transcriptional initiation events and promoter structure have classically been investigated in one or a few promoters, leading to general hypotheses of mechanisms for regulating gene activation. In recent years, however, the complete genomic DNA sequences have become available for an increasing number of organisms, providing a resource that has changed the scale and potential of researching transcriptional regulation. We now face the significant challenge of interpreting entire genomes of "simple" genetic code. Major projects are underway that employ these sequence data to annotate genomes at the functional level, in an effort to decipher the complex principles governing patterns of gene expression in eukaryotic organisms. For example, the ENCODE (Encyclopedia of DNA Elements) Consortium is utilizing multiple high-throughput biological and computational strategies to map every transcript and regulatory element in 30 Mb (1%) of the human genome, in preparation for expanding this study to the entire genome[13]. Such efforts are uncovering general features of gene regulation consistent with previous research, as well as revealing surprising new findings that support an increasingly complex and diverse view of promoter structure and function. Here, we review the progress toward a more complete understanding of transcriptional regulation at promoters, in light of recent genome-scale investigations.

1.2 Large-scale promoter discovery in eukaryotic genomes

Identification of transcriptional promoters throughout the genome is critical to increasing our understanding of their contributions to gene regulation. Because much can be learned from comparing multiple examples of these regulatory elements, efforts to curate our knowledge of promoter regions have been ongoing for over twenty years[14]. Rapidly improving high-throughput and bioinformatics approaches have accelerated the discovery and location of promoters and have enhanced the quality of characterization and annotation of these elements, but the goal has remained the same: to understand the mechanisms by which promoters regulate transcription. Numerous resources and techniques now contribute to the large-scale study and analysis of an ever-expanding library of eukaryotic promoters.

Because promoters are functionally and physically linked to the transcripts they generate, the completed sequencing of the human genome and the genomes of a growing number of other organisms has facilitated the use of sequence information of full-length transcripts to identify promoters involved in their regulation. Conventionally used to quantitatively monitor gene expression levels, transcript-capture techniques have been adapted to identify TSS with remarkable precision and genomic coverage. Several innovative strategies have been employed to collect large transcript-based sequence libraries. A modification of conventional cDNA cloning[15] enabled the precise capture of the sequence of the transcript 5’ end, and the adaptation of this strategy to large scale cDNA library construction[16] enabled the relatively streamlined assembly of a vast catalog of TSS. Recent updates to this Database of Transcription Start Sites (DBTSS) include expansion of human and mouse TSS data and the inclusion of additional organisms[17]. Similar technologies include Gene Identification Signature (GIS) analysis[18], 5’ end Serial Analysis of Gene Expression (5’ SAGE)[19,20], and Cap Analysis Gene Expression (CAGE)[21-23]. These advancements of conventional transcript analysis have made possible the high-throughput capture of 5’- and 3’-ends of entire transcriptomes in mouse and human systems. By matching the 5’ ends to genomic DNA sequences, it is possible to generate maps of putative promoter regions for known and novel genes that can be further characterized by various means.

In addition to transcript-based promoter identification, the maturation of technologies like ChIP-chip[24] has allowed the biochemical determination of promoters based on the protein-DNA interactions between the transcriptional machinery and the promoter sequences (see Figure 1.2). By examining genome-wide binding patterns of components of the PIC in human fibroblast cells, one study located over 10,000 active promoter sites and almost 1,200 novel promoters for previously unannotated transcriptional loci[25]. In addition to promoters for protein-coding genes, some of these novel promoters correspond to microRNA genes, whose transcripts were not amenable for identification by conventional cDNA cloning methods[26]. Therefore, the ChIP-chip approach complements the cDNA library-based method for promoter mapping.

Advances in bioinformatics also contribute to promoter discovery. While several general sequence features of promoters are known (discussed below in section III), the degeneracy and inconsistent presence of these sequences in promoters have long hindered the success of various computational approaches in identifying promoters on a genomic scale[27]. More recent efforts have integrated transcript data and multi-species sequence conservation information with first-exon-finding algorithms, offering a significant improvement in the accuracy of mammalian promoter identification[28]. Promoters identified in this study are curated in the Cold Spring Harbor Laboratory Mammalian Promoter Database (CSHLmpd), which also cross-references numerous established gene collections as well as promoters discovered in ChIP-chip and functional studies. The CSHLmpd is a useful complement to the Eukaryotic Promoter Database (EPD), which has grown exponentially from its original collection of 168 promoters[29] with the integration of numerous genome-scale data sets[14]. Additional valuable resources can be found in other public databases, including the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) and the UCSC Genome Browser (http://genome.ucsc.edu/)[30]. Both sites contain vast amounts of data from a variety of experimental and computational sources, as well as an array of powerful utilities for the visualization, analysis, and comparison of public and user data sets.

1.3 Signatures of promoters

The diverse approaches to promoter identification described above have provided unprecedented resources for large-scale promoter characterization. Recent advances in high-throughput experimental methods and computational analysis strategies have provided significant insight into the physical and functional features of promoters. One goal of such investigations is to define the "signature" of a promoter, that is, the sequence elements and chromatin features that dictate the promoter's regulatory properties (see Figure 1.3).

1.3.1 Sequence signatures

As noted above, the promoter consists of a core region immediately surrounding the TSS, and additional proximal promoter regions extending further upstream of the core promoter. Because the core promoter is the minimum region required for docking of the transcriptional machinery and initiation of basal transcription, extensive research in a variety of organisms has been devoted to uncovering the sequence motifs responsible for this critical step in gene regulation, revealing a collection of short regulatory DNA sequence elements conserved across species. While the first core promoter element has been known for almost thirty years, additional novel sequence elements have been discovered recently, emphasizing the importance of continued research of these regulatory sequences. Most of the canonical core promoter elements have been thoroughly reviewed elsewhere[2], but it is useful to describe their general features here (see Table 1.1) in light of recent genome-wide analyses of these elements. Note that there are no "universal" core promoter elements; the sequences described below are found in only a subset of promoters, and the origins and functional consequences of the resulting core promoter diversity are a topic of current study.

The first core promoter element identified was the TATA box, whose consensus sequence (TATAWAAR; degenerate nucleotides according to IUPAC code, http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html) was determined by comparison of 5' flanking regions in several organisms[31]. The TATA box is located approximately 25-30 bp upstream of the transcription start site in most eukaryotes, though in yeast it is found slightly further upstream[32]. It is typically recognized by the TATA binding protein (TBP) subunit of the general transcription factor TFIID[33], though additional related but distinct proteins can also recognize this element[34].

The initiator element (Inr; YYANWYY) immediately surrounds the transcription start site[35] and is found in promoters containing or lacking a TATA box. While the Inr can stimulate transcription independently of a TATA box, these two elements act synergistically when found together[36]. This element is recognized by the TAF1 and TAF2 subunits of TFIID[37].

The downstream promoter element (DPE; RGWYV)[38], is typically found in TATA-less promoters and functions with the Inr as a downstream counterpart to the TATA-box[39]. The DPE is located at +28 to +32 relative to the TSS, with this exact spacing critical to optimal transcription[40]. Like the TATA box and Inr, this element is recognized by TFIID, likely the TAF6 and TAF9 subunits, but not TBP[41]. There is evidence that the presence of a TATA-box or DPE in a promoter can influence its interactions with enhancers[42] and transcriptional activation or repression[43], suggesting multiple regulatory mechanisms acting at the core promoter.

The TFIIB recognition element (BRE; SSRCGCC) consists of the seven base pairs immediately upstream of the TATA box and, as its name suggests, it is bound by transcription factor IIB[44]. The BRE has been shown to both stimulate and repress transcriptional activity[45].

The motif ten element (MTE; CSARCSSAACGS) was identified in a computational survey of *Drosophila* promoters[46], located +18 to +29 downstream of the TSS and overlapping slightly with the 5’ end of the DPE. The MTE requires Inr and functions synergistically with the TATA-box or DPE, but can also function in a TATA- and DPE-independent manner and can compensate for mutations in either of these other elements[47]. It appears that the MTE contributes to interaction with TFIID.

Other core promoter motifs include the downstream core element (DCE)[48] and multiple start site downstream element (MED-1)[49], and continued research with an expanding library of well-annotated promoters has revealed additional putative regulatory motifs[50]. Another general sequence feature of many promoters in mammals is the presence of stretches of the CG dinucleotide, or “CpG islands”, which are underrepresented in the genome compared to what would be expected by chance for any given dinucleotide. Cytosines in DNA are often methylated to form 5-methyl cytosine (5mC), and the high frequency of spontaneous deamination of 5mC converts it to thymidine, resulting in the net loss of C at that position. Surviving CpG dinucleotides are therefore thought to be maintained by functional and evolutionary constraints for regulatory purposes. CpG island promoters typically lack a TATA-box[51], and the precise mechanisms of their core promoter function are not well understood.

Several recent large-scale analyses have confirmed the lack of universal core promoter elements, demonstrating that each element is found in subsets of promoters, with differing relative representation among species (see Table 1.1). For example, the TATA-box was once presumed to be a general feature of promoters, but genomic analyses clearly indicate that its presence is variable between species and actually atypical. A consensus TATA-box is present in only 33-43% of promoters in *Drosophila*[40,46], and in only about 10%-16% of mouse and human promoters[25,52-54]. Furthermore, while 69% of *Drosophila* promoters contain the Inr[40,46], only about 55% of human promoters possess this element[25]. In contrast, the DPE appears to be more abundant in human promoters (about 48%)[53] than in *Drosophila* (about 40%)[40]. CpG islands seem to be the most highly represented class of promoter element, with recent estimates of 79-88% of human promoters and 71% of mouse promoters[25,53] containing this feature, much higher than earlier estimates of about half of promoters[55]. Of course, these elements may be present in various combinations. For example, in *Drosophila* the TATA-box and DPE occur together in 14% of promoters[40], and 12% of TATA-containing vertebrate promoters also contain a BRE[44]. Further research is necessary to exhaustively catalog these and other core promoter elements and sequence variants throughout entire genomes, but the variety in core promoter structure within and between species suggests a significant role for core promoter diversity in transcriptional regulation, contrary to early single-gene studies that implied a universal promoter sequence.

1.3.2 Epigenetic signatures

Perhaps the most defining functional characteristic of an active promoter is the initiation of transcription at that promoter. Indeed, quantitative functional studies of human promoters demonstrate the expected strong correlation between promoter activity and endogenous transcript levels, confirming the promoter's key role in the rate of transcription[54]. But as static contributors to gene regulation, the presence or absence of core promoter elements is not informative as to the expression activity of the target transcript, and even transcript level is not always an accurate gauge of promoter activity due to various mechanisms of mRNA degradation or stabilization. Chromatin structure at promoters is recognized as an important determinant of gene expression[4], and the recent large-scale mapping of epigenetic features has revealed distinct chromatin signatures for active and inactive promoters. It is worth noting that classifying promoters as "active" or "inactive" simplifies a somewhat complicated situation. Promoter activity encompasses a continuum from weak expression to strong, and some chromatin features discussed below reflect that dynamic range of activity. Furthermore, some promoters might be maintained in a quasi-active state; these genes are not silenced by permanent repressive influences, yet the transcript originating from this promoter may not be actively expressed, perhaps waiting for a final regulatory event to initiate transcription. Such promoters can be distinguished from truly active or inactive promoters by referring to them as transcriptionally "competent." Active and competent promoters may share some features that are not present at inactive promoters, but it is worth noting that additional regulatory signals exist to elevate a promoter from competence to transcriptional activity.

Transcriptional regulatory events at promoters occur in the context of chromatin, which consists of ~146 bp of DNA wrapped around an octamer of histone proteins to form a nucleosome, resulting a repetitive and ordered structure originally viewed primarily as a means of DNA-packaging. However, we now know that the amino-terminal tails of the histones are subject to a wide variety of post-translational modifications [5] that influence the structure of the nucleosome and its interactions with DNA and regulatory proteins, including transcription factors, histone modifiers, chromatin remodelers, and the transcriptional machinery[56]. Variants of the histone proteins themselves also impact the nucleosome's structural and regulatory properties[57]. As it is generally understood that transcription factors are granted access to regulatory DNA sequences by permissive nucleosome conformations, local chromatin architecture (including histone modifications and nucleosome positioning) clearly plays a critical regulatory role at transcriptional promoters[4]. Here we examine some of the general features of chromatin associated with active promoters as revealed in recent genomic investigations in multiple organisms.

One key component of chromatin is, in fact, absent from active promoters: the nucleosome. Different experimental approaches in yeast have demonstrated depletion of nucleosomes at transcriptionally active promoters. ChIP-chip studies examining the enrichment patterns of core histones revealed a markedly reduced density of these proteins at the promoters of active genes genome-wide[58-61], indicating nucleosome depletion at these sites. High-resolution nucleosome mapping in yeast confirmed this observation, revealing a nucleosome free region (NFR) of ~150 bp in size located ~200 bp upstream of the start codon[62]. The nucleosomes flanking this NFR contain the histone variant H2A.Z[63-65], implicating H2A.Z in NFR formation or maintenance, though differences in experimental techniques make it unclear how H2A.Z enrichment relates to transcriptional activity. The significant structural differences between normal H2A and H2A.Z provide distinct protein interaction domains unique to this variant; these features may contribute to a role for H2A.Z in antagonizing gene silencing[66]. Interestingly, a short DNA sequence element was demonstrated to be responsible for NFR formation[64], consistent with the observation of sequence-dependent DNA-histone interactions in yeast promoter regions[61]; these findings further emphasize the connection between DNA sequence and chromatin structure.

Similar patterns of nucleosome depletion at active promoters were observed in *Drosophila*[67] and humans[68], contrary to an earlier study in mammalian cells[69] that found no change in nucleosome density at promoters. These recent findings are consistent with numerous reports demonstrating increased chromatin accessibility (as assayed by nuclease sensitivity) at promoters and other regulatory elements[70], and indicate that nucleosome depletion is an evolutionarily conserved mechanism of transcriptional regulation. An additional histone variant, H3.3, was found to be enriched at active promoters in *Drosophila*[67], further emphasizing the intimate relationship between nucleosome composition and transcriptional regulation. While the structure of H3.3 (and other H3 variants) is quite similar to that of normal H3, the recent "H3 barcode hypothesis"[71] proposes that subtle changes in nucleosome stability resulting from incorporation of H3 variants can influence protein interaction, nuclear localization, and post-translational modification, with profound impacts on gene regulation, epigenetic memory, and chromatin structuring.

The discovery that the histone proteins within nucleosomes could be covalently modified led to the proposal of the histone code hypothesis[72], wherein distinct functional and regulatory information is encoded in patterns of histone acetylation and methylation, among other possible modifications[5,56]. The field of epigenetics has exploded in recent years and it would be impossible to thoroughly cover it in this review, so we will focus on relevant global studies of histone modifications associated with gene activation (using current nomenclature[73]). As genomics technology has rapidly evolved over the past few years, so has the coverage, resolution, and specificity of the data gained from genome-wide epigenetic analyses. We will primarily discuss the most comprehensive current findings, acknowledging that they often confirm the results of many previous smaller-scale experiments. New and unexpected insights into promoter epigenetics have also been gained by the genome-wide expansion of previous single-gene findings.

Histone acetylation has long been found associated with active genetic regions, and many lysine residues within the various histone tails are subject to this modification[74]. Acetylation of histone lysines is a reversible modification controlled by two antagonistic protein families, the histone acetyltransferases (HATs) and histone deacetylases (HDACs). A genome-wide, high-resolution (~266 bp) assessment of histone acetylation in yeast revealed that acetylation of histone H3 lysines 9 (H3K9ac) and l4 (H3K14ac) and general acetylation of histone H4 (H4ac) are localized predominantly to promoters in a manner associated with transcriptional activity[60]. These modifications peak slightly downstream of the TSS. Similar ChIP-chip experiments in fly[75] and mammalian systems[25,69,76] demonstrated that acetylation of H3 and H4 is a conserved feature of transcriptionally active promoters. It is worth noting, however, that H3ac and H4ac have also been associated with some distal regulatory elements such as enhancers[68,76]. A single-nucleosome resolution study of residue-specific histone acetylation patterns in 500 kbp of the yeast genome offered additional insight, including the observation that specific lysines (H2AK7, H3K9, H3K14, H3K18, H4K5, H4K12) are hyperacetylated on nucleosomes at the 5’ ends of active genes, adjacent to a hypoacetylated region surrounding the active promoter;

intriguingly, principle component analysis revealed that the twelve histone modifications examined actually sort into two main classes (either promoter proximal or as a continuum through coding regions), rather than exhibiting independent distribution patterns[77]. Another study using histone-lysine mutants combined with global expression analysis suggested significant functional redundancy of residue-specific acetylation in histone H4, as only mutation of H4K16 caused specific changes in gene expression patterns[78]. These findings challenge the original hypothesis of a histone code with great combinatorial complexity conferred by distinct modifications, suggesting instead a simpler system in which multiple modifications play redundant roles in gene regulation, similar to the signaling network model of chromatin[79].

As with acetylation, histone lysines can be modified by methylation. Histone methylation seems more complex, however, as distinct histone methyltransferases (HMTs) can modify lysine residues by the addition of one, two, or three methyl groups, each of which appear to have distinct localization patterns and regulatory potential. Further, methylation is associated with both activation and repression of transcription, depending on the modified residue. Though lysine methylation was long thought to be irreversible, the recent discovery of histone demethylases[80] suggests that this modification may be as dynamic as acetylation. The aforementioned studies in yeast[60,77] revealed a gradient of methylation of H3K4 from 5' to 3' within actively transcribed genes, with tri-methylation of this residue (H3K4me3) peaking at the 5' end of the gene and giving way to di- and then mono-methylation (H3K4me2, H3K4me1) with increasing distance from the promoter. Like acetylation, these methylation patterns correlate with transcriptional activity, a relationship also generally observed in investigations of H3K4me3 and H3K4me2 in fly[75] and mammalian systems[25,69]. Another recent high-resolution study confirmed the H3K4me3-me2-me1 gradient at active human promoters[68]. H3K4me3 appears to mark active promoters exclusively, while H3K4me2 and H3K4me1 are also found elsewhere in the genome at other putative regulatory elements[68,69].

The chromatin features of inactive promoters are less well characterized, but the above studies demonstrated that inactive promoters generally lack the histone modifications associated with promoter activity, including acetylation of H3 and H4 and methylation of H3K4. Tri-methylation of H3K27 appears to be localized to promoters of repressed genes genome-wide[81,82]. Also, repressed genes are frequently located in heterochromatin[83], where the condensed structure ostensibly prevents transcription factor access to regulatory DNA sequences, though some characteristic features of open, active chromatin have been noted at inactive promoters in yeast[61,63,64].

In summary, genome-scale experiments in a variety of organisms from yeast to human indicate that transcriptionally active promoters are marked by nucleosome depletion, acetylation of several residues of H3 and H4 and tri-methylation of H3K4, and histone variants linked to transcription, while promoters of inactive genes generally lack these features. As noted, the majority of the histone modifications localize to the 5’ ends of genes, emphasizing the regulatory significance of the promoter region and hinting at a more simple histone code for promoters than originally thought. With the development of an ever-expanding repertoire of residue-specific antibodies and improvements in microarray and other high-throughput technologies, the next few years should see a wealth of high-resolution histone modification maps for the genomes of many organisms, which will be useful in decoding the regulatory mechanisms of histone modifications at promoters and other regulatory elements.

1.4 Promoter function and regulation

With the generation of large collections of promoters and the discovery of signature sequences and epigenetic features, many recent investigations have begun to examine the connections between DNA sequence, chromatin architecture, and promoter function, providing insight into the molecular mechanisms of transcriptional regulation at promoters. Preliminary regulatory networks were often assembled on the basis of transcript expression analysis, whereby groups of co-expressed genes were postulated to share common control circuits. This method, while a useful starting point, cannot distinguish between direct and indirect regulatory targets. To actually decipher the regulatory code underlying co-regulated genes, the expression patterns must be supplemented with knowledge of the regulatory proteins and epigenetic features present at the promoters of active and inactive genes. Several strategies, such as ChIP-chip, are currently employed to determine the direct targets of a variety of transcriptional regulators[24].

1.4.1 Regulatory networks

Sequence-specific transcription factors (TFs) play a critical role in regulating transcription by recruiting coactivators and promoting the formation of the PIC[9,84]. Consequently, many investigations have focused on the discovery of direct targets of TF binding. TF consensus binding motifs are often somewhat degenerate, causing sequence-based computational methods to predict many thousands of binding sites for a given TF, only a fraction of which may be biologically relevant. Indeed, even binding sites for which the cognate TF has a very high affinity in vitro are not necessarily bound in vivo, consistent with our understanding of mechanisms underlying tissue-specific programs of gene expression. Conversely, TF targets may not contain consensus binding motifs[85,86], suggesting that the TFs are binding to uncharacterized motifs or through cooperation with additional factors. Thus, any apparent connection between expression data and promoter DNA sequence is, at best, circumstantial evidence of TF binding.

The development of technologies like ChIP-chip enabled the rapid and direct biochemical purification of DNA sequences bound by TFs in the genome in vivo and the subsequent generation of target maps and transcriptional regulatory networks. The first global studies of TF binding in yeast revealed that, in spite of the presence of consensus binding motifs for Gal4 and Ste12 throughout the yeast genome, these factors localize to the promoters of functionally related genes to form distinct regulatory modules[87]. Similar patterns were observed for the TF Rap1[88], suggesting that additional features such as chromatin architecture are involved in the selective binding of these TFs to promoters. Extension of this assay to over 100 yeast TFs revealed that many yeast promoters are bound by multiple TFs, echoing the combinatorial complexity postulated for higher eukaryotes[89]. This study also introduced the integration of network motifs (such as autoregulation, feedforward, and multi-input) with expression data to construct regulatory networks for processes like metabolism and the cell cycle. Such strategies were expanded to include over 200 yeast TFs[90], resulting in the discovery of novel regulatory DNA sequences, insights into promoter structure, and a system of TF classification based on functional binding data. These experiments provided the first broad view of promoter topography on a genomic scale.

Such investigations are more complex in higher eukaryotes. Metazoans are composed of many different cell types, requiring a much larger arsenal of TFs to regulate elaborate patterns of differentiation, homeostasis, and environmental response, not to mention the corresponding increase in the size and complexity of the genome. Given the larger size of mammalian genomes, initial location analyses in mouse and human systems examined patterns of TF binding using microarrays representing thousands of promoter regions, which at the time were the only regulatory elements that could be effectively located. Even examining these small fractions of the genome in tissue-specific contexts proved enlightening. For example, an investigation of TCF4 target genes revealed that the EDN1 oncogene is a direct regulatory target of β-catenin in colon cancer, providing important insight into the activation of this growth factor in colon and other cancers[91]. An examination of several myogenic TFs at promoters in proliferating and differentiating mouse myoblasts uncovered a complex, dynamic network governing skeletal myogenesis as well as unexpected involvement in stress response and regeneration[92]. Studying the binding patterns of HNFs in human liver and pancreatic cells revealed distinct and common regulatory targets between tissues and provided mechanistic insight into the potential of HNF4α misregulation to contribute to type II diabetes[93]. Similar experiments with c-Myc and its binding partner Max in Burkitt's lymphoma cells revealed that over 15% of the promoters studied are bound by both factors[85], comparable to observations in HL60 cells[94]. The surprisingly large number of targets for these TFs suggested a general role for c-Myc in global transcriptional regulation, a model supported by additional experiments analyzing Myc targets and gene expression in *Drosophila*[95].

While these studies provoked new ideas about transcriptional regulation at mammalian promoters, the coverage and resolution of the microarray platforms used in these experiments limited the insight that could be gained. Improved genome sequence annotation and technological advances in microarray synthesis and analysis led to the development of "tiling" arrays, wherein short oligonucleotide probes provide continuous coverage along large regions of the genome, in contrast to previous arrays that sampled isolated chunks of promoters or other genomic sites. A more advanced promoter microarray platform was developed that covered 10-kb regions tiling almost 18,000 human promoters with 60-mer oligos, and used to identify targets of Oct4, Sox2, and Nanog in human embryonic stem cells (hESC)[96]. These experiments yielded precise binding sites of these TFs within their target promoters and revealed a large number of targets common to all three factors, forming coordinated feed-forward and auto-regulatory loops with intriguing implications in hESC pluripotency and self-renewal.

In addition to their utility in finding regulatory targets of TFs, tiling arrays have enabled the unbiased discovery of regulatory regions through analysis of genomic binding patterns of TF and other proteins (most of the chromatin architecture discussed in section III was determined using tiling arrays). Some TFs are found primarily at promoters, like YY1 [K. Wang and B. Ren, unpublished data] and E2F1[97], and their binding with RNAPII at many promoters (>20% for E2F1) suggests a general role for these TFs in transcriptional regulation. Interestingly, however, a growing number of experiments show that many TFs bind to distal sites throughout the genome, far from any annotated genes. Tiling arrays covering human chromosomes 21 and 22 revealed that only a small fraction of p53 binding occurred near known promoters[98], and similar patterns have been observed for estrogen receptor (ER) in the same regions[99]; NF-κB, CREB, and STAT1/2 on chromosome 22[100-102]; and p53 throughout the entire human genome[103]. The widespread binding patterns of these TFs are reminiscent of the genomic distribution of distal regulatory elements like enhancers, and several lines of experimental evidence support a physiological enhancer function for the distal ER binding sites[99]. Another explanation proposed for promoter-distal binding involves regulation of non-coding RNAs[98]. Distal binding sites aside, these experiments identified many novel target genes for these TFs and provided insight into the requirement for and sequence of consensus binding motifs. Additionally, the overlap of TF binding at promoters observed within the experiments above lends support to theories of a combinatorial code in transcriptional regulation in higher eukaryotes, wherein the coordinated action of several TFs at a given promoter is required for precise regulation of expression. Further assessment of binding patterns of additional TFs in multiple tissues will hopefully lead to the development of complete human transcriptional regulatory networks that address the complex genetic mechanisms underlying development and disease.

1.4.2 Regulatory mechanisms

In addition to identifying targets of specific TFs, location analysis of components of the basal transcriptional machinery has provided some insight into general mechanisms of gene regulation. The majority of active promoters in human fibroblasts are bound by the general TF TAF1[25], consistent with the critical role of this protein in PIC assembly. It has also been demonstrated that hypophosphorylated RNAPII is localized primarily at promoters in humans, while total RNAPII is found enriched throughout genes, primarily at exons[25,104]. These findings are consistent with existing models of transcriptional initiation control through regulated phosphorylation of RNAPII[6], and support coordinated mechanisms for transcriptional elongation and mRNA processing events. About 75% of promoters occupied by the PIC appeared to be transcriptionally active, indicating that TAF1 and RNAPII occupancy are a general feature of active promoters, even considering the diversity of core promoter elements found in these promoters[25]. Promoters marked by a PIC but with no evidence of transcription could reflect the competent promoters mentioned earlier, awaiting further activating signals. It is important to remember, however, that the basal transcriptional machinery is not always composed of the same subunits[105], so further large-scale experiments are needed to determine the precise constitution of the PIC at diverse promoters. Additional TAF1 and RNAPII binding distal to known promoters may signify the presence of novel promoters or other putative regulatory elements, providing some insight into mechanisms of interaction between promoters and distal elements like enhancers. Comparison of these sites to high-resolution maps of histone modifications and TF binding should prove informative.

Owing to the diversity of sequence-specific transcription factors in eukaryotic genomes and the coactivators through which they mediate transcriptional regulation[106] and considering the tissue-specificity of many gene expression patterns, promoter activation is difficult to generalize at the level of the sequence-specific TF. Some common patterns of coregulator localization, however, have recently begun to emerge. Most active promoters in yeast are occupied by HAT enzymes like Gcn5 and Esa1[60,107], consistent with models linking gene activation to acetylation of histones by these enzymes and with the acetylation patterns observed at active promoters as discussed in section III. Similarly, the HAT p300 has been observed at many active promoters in human cells[68], supporting a conserved role for such factors in positively regulating transcription. The precise purpose of histone acetylation at promoters is not yet known, but several lines of thought address the mechanistic significance of this modification. Many transcriptional regulatory proteins (including TAF1) possess bromodomains capable of recognizing acetylated lysines, which would serve to initiate and stabilize interactions between these proteins and the promoter region[108]. Histone acetylation also appears to influence binding of sequence specific transcription factors to DNA by revealing some consensus binding motifs and occluding others [J. Lanier and E. Turner, personal communication], similar to the formation of the NFR that presumably facilitates binding of the transcriptional machinery. Furthermore, histone deacetylation has been linked to transcriptional elongation[12], so it is possible that the relatively hyperacetylated histones at promoters serve to distinguish physically adjacent yet functionally discrete components of a genetic unit.

As with the various HATs, the HMT responsible for catalyzing the tri-methylation of H3K4 in yeast, Set1, has also been demonstrated to associate with the promoter regions of active genes[109], and similar patterns were observed with the human Set1 homolog, MLL1[110]. Again, the functional significance of this modification has yet to be entirely deciphered, but as with acetylation, methlyated lysines can be recognized by numerous regulatory proteins that contain chromodomains[108,111]. Additional evidence suggests that H3K4me3 may be involved in regulating HAT and HDAC activity in the rapid turnover of acetylation at active promoters[112]. The hyperacetylation could then be preferentially maintained at promoters while H3K4me2, H3K4me1, and/or other distinct methylated histone residues facilitate the aforementioned deacetylation that occurs in coding regions, again creating a functional compartmentalization mediated and marked by a methylation gradient. Intriguingly, several recent reports identify PHD-finger-containing proteins as novel recognizers of H3K4me3 with implications in both maintenance and repression of gene expression[113-116], suggesting that H3K4me3 is a multi-purpose marker for active promoters, recognized in specific contexts by activator or repressor proteins in response to cellular signaling pathways. Further experiments are required to more finely resolve these regulatory mechanisms, but the presence of various HATs and HMTs at the majority of active promoters is consistent with a general role for these factors in transcriptional regulation.

One attribute common to histone features and transcription factor binding at promoters is their association with maintaining patterns of gene activity through mitosis[117], even when these promoters are not transcriptionally active. This "gene bookmarking" supports the concept of a cellular memory, in which epigenetic features associated with gene activity persist through transcriptional inactivation to mark these genes for potential subsequent reactivation, protecting them from permanent silencing through incorporation into heterochromatin. Additional genome-scale studies will be useful in elucidating the connections between transcriptional activation and maintenance of promoter competence and activity.

1.4.3 Connecting sequence to regulation

Recent investigations have begun to reveal more of the relationship between sequence features of promoters and their function and regulation. Comparative computational analysis of a large number of human, rodent, and dog promoters uncovered a variety of conserved DNA sequences, including most known TF consensus motifs and many novel putative regulatory sequences[50]. The validity of the novel sequences is supported by several lines of evidence, including motif enrichment in tissue-specific promoters, conserved positional preference, and the clustering of motif copies within promoters. Whether or not these sequences represent novel binding motifs for TFs or are even truly functional in vivo has yet to be determined, but comparisons of these findings with high-resolution maps of TF binding and histone modifications will likely yield valuable insight into the sequences underlying protein-DNA interactions.

Established core promoter features are also connected to gene regulatory and functional properties. CpG island promoters are generally associated with ubiquitously expressed housekeeping genes, while TATA-box promoters appear to be more tightly and specifically regulated[23], in support of previous findings. This trend also translates to the precision of transcriptional initiation from these classes of promoters; in contrast to more defined TSS in TATA-box promoters, multiple TSS spanning upwards of 100 bp are often detected in CpG promoters, most recently shown on a genomic scale by Carninci et al[23]. Consistent with expression-based observations, a functional analysis of hundreds of putative promoters in 16 human cells lines showed that 86% of promoters exhibiting ubiquitous strong activity in all cell lines overlapped CpG islands[54]. Further division of mammalian promoters into four classes based on CG content upstream and downstream of the TSS revealed connections between different CG enrichment patterns and core promoter elements, expression, and gene function, with potential differences between mouse and human promoters including variable representation of certain core promoter elements[53].

Additional evidence links CpG islands to bidirectional promoters, which represent over 10% of human promoters; intriguingly, 77% of bidirectional promoters are located within CpG islands while only 8% of these promoters contain a TATA-box[118]. This study also found conservation of these bidirectional promoter structures in mouse, and uncovered interesting relationships between promoter bidirectionality and gene function and regulation of expression. While this investigation showed that a significant proportion of genes appear to share promoter sequences, other recent studies have revealed widespread usage of alternative promoters throughout mammalian genomes by examining binding of the transcriptional machinery to multiple sites at gene 5' ends[25], transcript-based identification of adjacent but distinct TSS[23,119], and functional analysis of putative promoters[54]. In addition to demonstrating the tissue-specificity of many promoters even without the influence of distal regulatory elements, this functional study also found distinct regions of the proximal promoter that are related to transcriptional activity, including the intriguing general presence of positive regulatory regions 40-350 bp upstream of the TSS and negative regulatory regions 350-1000 bp upstream of the TSS[119]. The mechanisms of regulation by these regions have yet to be determined, but such findings clearly highlight the importance of considering the proximal promoter when studying transcriptional activation and repression. In addition to providing insight into the general functional properties of promoters, such large-scale functional assays also form the basis for investigating the contributions of DNA sequence and chromatin structure to tissue-specific gene expression and promoter usage.

1.5 Conclusion and perspectives

Constantly evolving computational and experimental methodologies will continue to make significant contributions to our knowledge of promoter signatures at the DNA sequence and epigenetic levels. Genomic sequencing of additional organisms and advances in sequence alignment strategies will provide expanded resources for comparative promoter analyses, potentially revealing novel promoter sequence elements with transcriptional regulatory properties. Furthermore, only a small fraction of the >100 known histone modifications have currently been mapped on a large scale. Future studies will investigate these modifications in other systems and will expand to include additional modifications and histone variants, contributing to a more complete understanding of the chromatin architecture at promoters and other transcriptional regulatory elements. Another current focus of epigenetic research is examining the patterns of DNA-methylation, wherein methylation of cytosine (usually within CpG islands) represses gene expression by inducing heterochromatin formation or by interfering with transcription factor binding[120]. The recent development of a large-scale DNA-methylation profiling assay enabled the generation of a DNA-methylation map of the entire human genome[121], revealing surprising results related to the role of DNA-methylation in heterochromatin formation, X chromosome silencing, and development of malignant cancer. A similar study examining a large collection of human promoters uncovered evidence for a targeted instructive mechanism for DNA-methylation of promoters in cancer cells[122]. Additional experiments are needed to resolve the mechanisms underlying DNA-methylation during development and oncogenesis and the impact of this modification on transcriptional regulation.

Significant progress has been made in locating promoters throughout the genome, identifying signature features of their DNA sequence and chromatin architecture, and describing some of the regulatory proteins present at these sites, but much work remains to unravel the precise mechanisms by which active promoter structures are generated, regulated, and dismantled. To complement the considerable insight gained by analyzing evolutionary conservation of DNA sequence, additional research must identify all proteins involved in transcription, reveal the extent to which the regulatory structures and mechanisms of promoters are conserved across species, and relate the consequences of diverging structure and function to species-specific transcriptional regulation programs. Improvement of existing genomic strategies and the development of novel approaches will solve the complex regulatory code of eukaryotic transcriptional promoters, opening new doorways to understanding human disease, development, and evolution.

1.6 Acknowledgment

Chapter 1, in full, is a reprint of the material as it appears in “The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome” in Cellular and Molecular Life Sciences 2007, Heintzman ND and Ren B. The dissertation author was the primary investigator and author of this publication.

1.7 Tables and Figures

Table 1.1: Summary of sequence and frequency of core promoter elements

Core element	**Position relative to TSS***	**Consensus Sequence****	**Frequency in promoters**	
			Flies	**Vertebrates**
TATA	approx. -31 to -26	TATAWAAR	33-43%	10-16%
Inr	-2 to +4	YYANWYY	69%	55%
DPE	+28 to +32	RGWYV	40%	48%
BRE	approx. -37 to -32	SSRCGCC	-	12-62%
MTE	+18 to +29	CSARCSSAACGS	8.5%	-

* the TSS is assigned to position +1

** degenerate nucleotides represented using IUPAC codes

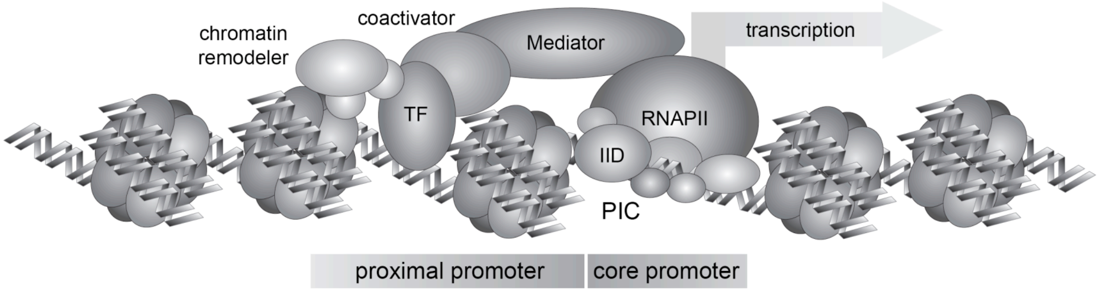

Figure 1.1: Typical structure of an active eukaryotic promoter.
The promoter consists of a core promoter region immediately surrounding the transcription start site, adjacent to a more extended proximal promoter region. RNA polymerase II (RNAPII) and various general transcription factors (for example, transcription factor IID) form the pre-initiation complex (PIC) around the transcriptional start site. Other transcriptional regulatory proteins including Mediator, chromatin remodelers, coactivators, and sequence-specific transcription factors (TF) are involved in regulating transcription at the promoter. All of these events occur in the context of chromatin, made up of DNA wrapped around octamers of histone proteins.

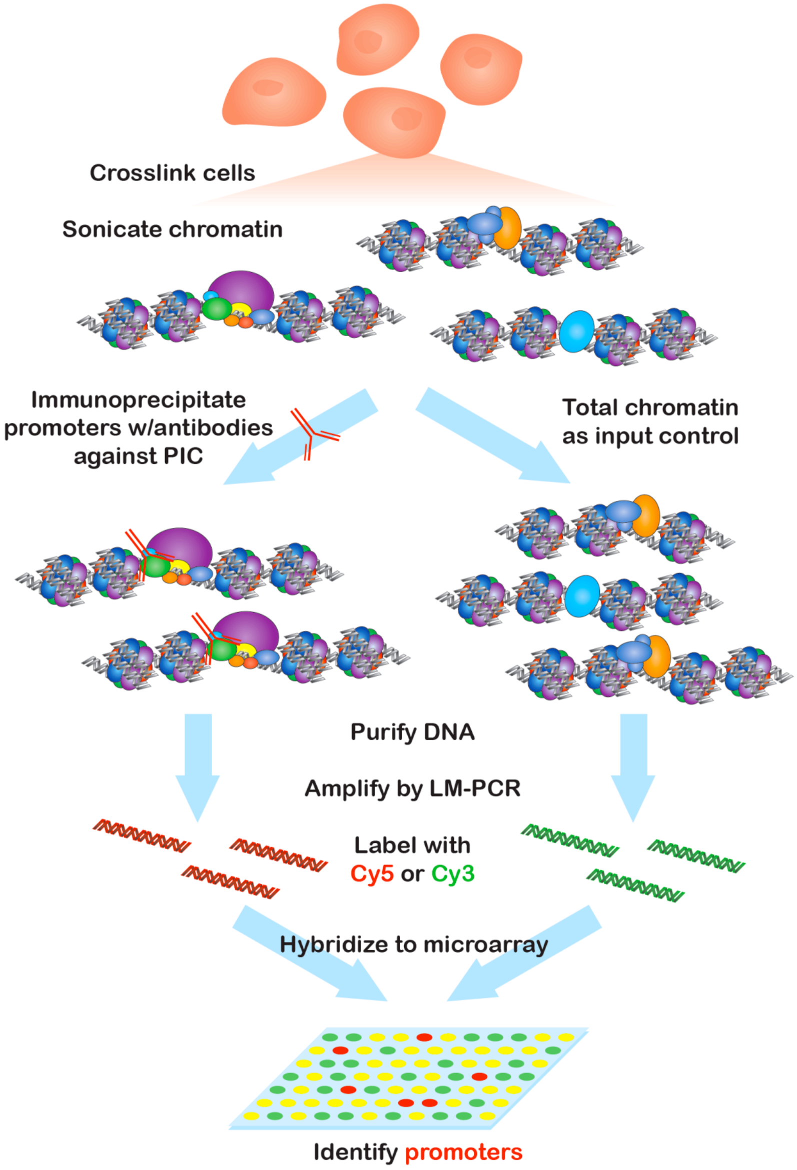

Figure 1.2: Promoter discovery using ChIP-chip.
Cells are treated with formaldehyde to chemically crosslink DNA and interacting proteins. Chromatin is isolated and sheared to small pieces by sonication, then subjected to immunoprecipitation with antibodies specific to components of the transcriptional machinery, in this case TFIID. Promoter fragments bound by TFIID will be enriched in the IP sample relative to a total chromatin control sample. DNA from both samples is purified, amplified, and labeled with fluorescent dye, then hybridized to a microarray covering large continuous stretches of the human genome. Promoters are identified on the basis of their enrichment in the IP sample, visualized as a red spot on the microarray.

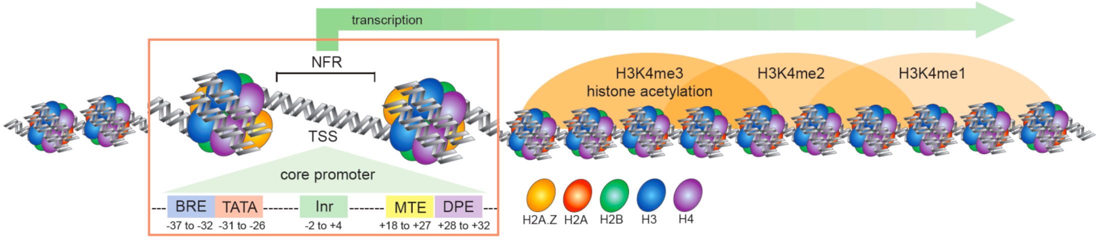

Figure 1.3: Signatures of active promoters.
A nucleosome free region (NFR) surrounds the transcriptional start site (TSS) in the core promoter, which may contain core promoter elements including BRE, TATA, Inr, MTE, DPE, or others (positions are relative to the +1 TSS within the Inr; please see detailed explanation of these elements in the main text and in Table 1.1). The nucleosomes flanking the NFR contain the histone variant H2A.Z, while other nucleosomes contain normal H2A and other histone proteins that are subject to various modifications. Histone acetylation peaks just downstream of the promoter, while methylation of histone 3 lysine 4 is present in a gradient, from tri-methylation (H3K4me3) at the promoter, to di- and then mono-methylation (H3K4me2, H3K4me1) with increasing distance from the promoter into the transcribed region. This diagram is a composite of features determined in yeast, fly, and mammalian systems; it is representative of some important characteristics of promoters identified in large-scale studies.

1.8 References

1. Orphanides G, Reinberg D. A unified theory of gene expression. *Cell* 2002;108:439-51.

2. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. *Annu Rev Biochem* 2003;72:449-79.

3. West AG, Fraser P. Remote control of gene transcription. *Hum Mol Genet* 2005;14 Spec No 1:R101-11.

4. Mellor J. The dynamics of chromatin remodeling at promoters. *Mol Cell* 2005;19:147-57.

5. Nightingale KP, O'Neill L P, Turner BM. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr Opin Genet Dev* 2006;16:125-36.

6. Hahn S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* 2004;11:394-403.

7. Butler JE, Kadonaga JT. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* 2002;16:2583-92.

8. Glass CK, Rosenfeld MG. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev* 2000;14:121-41.

9. Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 2000;14:2551-69.

10. Lewis BA, Reinberg D. The mediator coactivator complex: functional and physical roles in transcriptional regulation. *J Cell Sci* 2003;116:3667-75.

11. Ju BG, Lunyak VV, Perissi V, Garcia-Bassets I, Rose DW, Glass CK, Rosenfeld MG. A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. *Science* 2006;312:1798-802.

12. Lieb JD, Clarke ND. Control of transcription through intragenic patterns of nucleosome composition. *Cell* 2005;123:1187-90.

13. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306:636-40.

14. Schmid CD, Perier R, Praz V, Bucher P. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* 2006;34:D82-5.

15. Maruyama K, Sugano S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* 1994;138:171-4.

16. Suzuki Y, Yoshitomo-Nakagawa K, Maruyama K, Suyama A, Sugano S. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* 1997;200:149-56.

17. Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S. DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Res* 2006;34:D86-9.

18. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2005;2:105-11.

19. Hashimoto S, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K. 5'-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol* 2004;22:1146-9.

20. Kasai Y, Hashimoto S, Yamada T, Sese J, Sugano S, Matsushima K, Morishita S. 5'SAGE: 5'-end Serial Analysis of Gene Expression database. *Nucleic Acids Res* 2005;33:D550-2.

21. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 2003;100:15776-81.

22. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S. The transcriptional landscape of the mammalian genome. *Science* 2005;309:1559-63.

23. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;38:626-635.

24. Kim TH, Ren B. Genome-wide Analysis of Protein-DNA Interactions. *Annual Review of Genomics and Human Genetics* 2006;7.

25. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. *Nature* 2005;436:876-80.

26. Carthew RW. Gene regulation by microRNAs. *Curr Opin Genet Dev* 2006;16:203-8.

27. Bajic VB, Tan SL, Suzuki Y, Sugano S. Promoter prediction analysis on the whole human genome. *Nat Biotechnol* 2004;22:1467-73.

28. Xuan Z, Zhao F, Wang J, Chen G, Zhang MQ. Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biol* 2005;6:R72.

29. Bucher P, Trifonov EN. Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res* 1986;14:10009-26.

30. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res* 2002;12:996-1006.

31. Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem* 1981;50:349-83.

32. Struhl K. Yeast transcriptional regulatory mechanisms. *Annu Rev Genet* 1995;29:651-74.

33. Burley SK, Roeder RG. Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem* 1996;65:769-99.

34. Berk AJ. TBP-like factors come into focus. *Cell* 2000;103:5-8.

35. Smale ST, Baltimore D. The "initiator" as a transcription control element. *Cell* 1989;57:103-13.

36. Smale ST, Schmidt MC, Berk AJ, Baltimore D. Transcriptional activation by Sp1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID. *Proc Natl Acad Sci U S A* 1990;87:4509-13.

37. Chalkley GE, Verrijzer CP. DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. *Embo J* 1999;18:4835-45.

38. Kadonaga JT. The DPE, a core promoter element for transcription by RNA polymerase II. *Exp Mol Med* 2002;34:259-64.

39. Burke TW, Kadonaga JT. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* 1996;10:711-24.

40. Kutach AK, Kadonaga JT. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol Cell Biol* 2000;20:4754-64.

41. Burke TW, Kadonaga JT. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev* 1997;11:3020-31.

42. Butler JE, Kadonaga JT. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* 2001;15:2515-9.

43. Willy PJ, Kobayashi R, Kadonaga JT. A basal transcription factor that activates or represses transcription. *Science* 2000;290:982-5.

44. Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* 1998;12:34-44.

45. Evans R, Fairley JA, Roberts SG. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes Dev* 2001;15:2945-9.

46. Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* 2002;3:RESEARCH0087.

47. Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* 2004;18:1606-17.

48. Lewis BA, Kim TK, Orkin SH. A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc Natl Acad Sci U S A* 2000;97:7172-7.

49. Ince TA, Scotto KW. Differential utilization of multiple transcription start points accompanies the overexpression of the P-glycoprotein-encoding gene in Chinese hamster lung cells. *Gene* 1995;156:287-90.

50. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 2005;434:338-45.

51. Blake MC, Jambou RC, Swick AG, Kahn JW, Azizkhan JC. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol Cell Biol* 1990;10:6632-41.

52. Bajic VB, Choudhary V, Hock CK. Content analysis of the core promoter region of human genes. *In Silico Biol* 2004;4:109-25.

53. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C, Kawai J, Hume DA, Carninci P, Hayashizaki Y. Mice and men: their promoter properties. *PLoS Genet* 2006;2:e54.

54. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res* 2006;16:1-10.

55. Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res* 2001;11:677-84.

56. Margueron R, Trojer P, Reinberg D. The key to development: interpreting the histone code? *Curr Opin Genet Dev* 2005;15:163-76.

57. Sarma K, Reinberg D. Histone variants meet their match. *Nat Rev Mol Cell Biol* 2005;6:139-49.

58. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. Global nucleosome occupancy in yeast. *Genome Biol* 2004;5:R62.

59. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* 2004;36:900-5.

60. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young

RA. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 2005;122:517-27.

61. Sekinger EA, Moqtaderi Z, Struhl K. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell* 2005;18:735-48.

62. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* 2005;309:626-30.

63. Guillemette B, Bataille AR, Gevry N, Adam M, Blanchette M, Robert F, Gaudreau L. Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol* 2005;3:e384.

64. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD. Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell* 2005;123:233-48.

65. Zhang H, Roberts DN, Cairns BR. Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. *Cell* 2005;123:219-31.

66. Raisner RM, Madhani HD. Patterning chromatin: form and function for H2A.Z variant nucleosomes. *Curr Opin Genet Dev* 2006;16:119-24.

67. Mito Y, Henikoff JG, Henikoff S. Genome-scale profiling of histone H3.3 replacement patterns. *Nat Genet* 2005;37:1090-7.

68. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007;39:311-8.

69. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR, Schreiber SL, Lander ES. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 2005;120:169-81.

70. Felsenfeld G. Chromatin unfolds. *Cell* 1996;86:13-9.

71. Hake SB, Allis CD. Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis". *Proc Natl Acad Sci U S A* 2006;103:6428-35.

72. Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000;403:41-5.

73. Turner BM. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol* 2005;12:110-2.

74. Turner BM. Decoding the nucleosome. *Cell* 1993;75:5-8.

75. Schubeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, van Leeuwen F, Gottschling DE, O'Neill LP, Turner BM, Delrow J, Bell SP, Groudine M. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* 2004;18:1263-71.

76. Roh TY, Cuddapah S, Zhao K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* 2005;19:542-52.

77. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ. Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* 2005;3:e328.

78. Dion MF, Altschuler SJ, Wu LF, Rando OJ. Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A* 2005;102:5501-6.

79. Schreiber SL, Bernstein BE. Signaling network model of chromatin. *Cell* 2002;111:771-8.

80. Bannister AJ, Kouzarides T. Reversing histone methylation. *Nature* 2005;436:1103-6.

81. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, Bell GW, Otte AP, Vidal M, Gifford DK, Young RA, Jaenisch R. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 2006;441:349-53.

82. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 2006;125:301-13.

83. Craig JM. Heterochromatin--many flavours, common themes. *Bioessays* 2005;27:17-28.

84. Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 2004;116:247-57.

85. Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc Natl Acad Sci U S A* 2003;100:8164-9.

86. Zhang X, Odom DT, Koo SH, Conkright MD, Canettieri G, Best J, Chen H, Jenner R, Herbolsheimer E, Jacobsen E, Kadam S, Ecker JR, Emerson B, Hogenesch JB, Unterman T, Young RA, Montminy M. Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. *Proc Natl Acad Sci U S A* 2005;102:4459-64.

87. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. *Science* 2000;290:2306-9.

88. Lieb JD, Liu X, Botstein D, Brown PO. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* 2001;28:327-34.

89. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002;298:799-804.

90. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature* 2004;431:99-104.

91. Kim TH, Xiong H, Zhang Z, Ren B. beta-Catenin activates the growth factor endothelin-1 in colon cancer cells. *Oncogene* 2005;24:597-604.

92. Blais A, Tsikitis M, Acosta-Alvear D, Sharan R, Kluger Y, Dynlacht BD. An initial blueprint for myogenic differentiation. *Genes Dev* 2005;19:553-69.

93. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA. Control of pancreas and liver gene expression by HNF transcription factors. *Science* 2004;303:1378-81.

94. Mao DY, Watson JD, Yan PS, Barsyte-Lovejoy D, Khosravi F, Wong WW, Farnham PJ, Huang TH, Penn LZ. Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr Biol* 2003;13:882-6.

95. Orian A, van Steensel B, Delrow J, Bussemaker HJ, Li L, Sawado T, Williams E, Loo LW, Cowley SM, Yost C, Pierce S, Edgar BA, Parkhurst SM, Eisenman RN. Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. *Genes Dev* 2003;17:1101-14.

96. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 2005;122:947-56.

97. Bieda M, Xu X, Singer MA, Green R, Farnham PJ. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* 2006;16:595-605.

98. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 2004;116:499-509.

99. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 2005;122:33-43.

100. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A* 2003;100:12247-52.

101. Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, Weissman S, Snyder M. CREB binds to multiple loci on human chromosome 22. *Mol Cell Biol* 2004;24:3804-14.

102. Hartman SE, Bertone P, Nath AK, Royce TE, Gerstein M, Weissman S, Snyder M. Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes Dev* 2005;19:2953-68.

103. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y. A global map of p53 transcription-factor binding sites in the human genome. *Cell* 2006;124:207-19.

104. Brodsky AS, Meyer CA, Swinburne IA, Hall G, Keenan BJ, Liu XS, Fox EA, Silver PA. Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol* 2005;6:R64.

105. Hochheimer A, Tjian R. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes Dev* 2003;17:1309-20.

106. Taatjes DJ, Marr MT, Tjian R. Regulatory diversity among metazoan co-activator complexes. *Nat Rev Mol Cell Biol* 2004;5:403-10.

107. Robert F, Pokholok DK, Hannett NM, Rinaldi NJ, Chandy M, Rolfe A, Workman JL, Gifford DK, Young RA. Global position and recruitment of HATs and HDACs in the yeast genome. *Mol Cell* 2004;16:199-209.

108. de la Cruz X, Lois S, Sanchez-Molina S, Martinez-Balbas MA. Do protein motifs read the histone code? *Bioessays* 2005;27:164-75.

109. Ng HH, Robert F, Young RA, Struhl K. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* 2003;11:709-19.

110. Guenther MG, Jenner RG, Chevalier B, Nakamura T, Croce CM, Canaani E, Young RA. Global and Hox-specific roles for the MLL1 methyltransferase. *Proc Natl Acad Sci U S A* 2005;102:8603-8.

111. Daniel JA, Pray-Grant MG, Grant PA. Effector proteins for methylated histones: an expanding family. *Cell Cycle* 2005;4:919-26.

112. Hazzalin CA, Mahadevan LC. Dynamic acetylation of all lysine 4-methylated histone H3 in the mouse nucleus: analysis at c-fos and c-jun. *PLoS Biol* 2005;3:e393.

113. Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD, Patel DJ. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* 2006.

114. Pena PV, Davrazou F, Shi X, Walter KL, Verkhusha VV, Gozani O, Zhao R, Kutateladze TG. Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* 2006.

115. Shi X, Hong T, Walter KL, Ewalt M, Michishita E, Hung T, Carney D, Pena P, Lan F, Kaadige MR, Lacoste N, Cayrou C, Davrazou F, Saha A, Cairns BR, Ayer DE, Kutateladze TG, Shi Y, Cote J, Chua KF, Gozani O. ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* 2006.

116. Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P, Wu C, Allis CD. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* 2006.

117. Sarge KD, Park-Sarge OK. Gene bookmarking: keeping the pages open. *Trends Biochem Sci* 2005;30:605-10.

118. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. An abundance of bidirectional promoters in the human genome. *Genome Res* 2004;14:62-6.

119. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, Ishii S, Sugiyama T, Saito K, Isono Y, Irie R, Kushida N, Yoneyama T, Otsuka R, Kanda K, Yokoi T, Kondo H, Wagatsuma M, Murakawa K, Ishida S, Ishibashi T, Takahashi-Fujii A, Tanase T, Nagai K, Kikuchi H, Nakai K, Isogai T, Sugano S. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* 2006;16:55-65.

120. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002;16:6-21.

121. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 2005;37:853-62.

122. Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, Segal E, Pikarski E, Young RA, Niveleau A, Cedar H, Simon I. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet* 2006;38:149-53.

Chapter 2

Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome

This published chapter includes supplementary figures, supplementary tables, and supplementary information and data files, all of which are available at GEO (accession number GSE6273), at http://licr-renlab.ucsd.edu/download.html, and/or at the UCSC genome browser http://genome.ucsc.edu

Abstract

We report here that human promoters and enhancers are associated with distinct chromatin signatures that can be employed to predict these classes of regulatory elements in the human genome. Using a combination of chromatin immunoprecipitation (ChIP) and microarray (ChIP-chip) experiments, we generated high-resolution maps of the chromatin architecture along 30 Mbp of the human genome, located promoters and enhancers in these regions, and characterized the

histone modification features of these regulatory elements. We found that active promoters are marked by tri-methylation of histone H3 lysine 4 (H3K4me3) at the transcriptional start site (TSS) while enhancers are marked by H3K4me1 but not H3K4me3. We developed computational prediction algorithms that employ the distinct chromatin signatures to identify new promoters and enhancers, predicting over 200 promoters and 400 enhancers within the 30 Mbp regions. This approach correctly predicted over 84% of the regulatory elements identified in an independent, unbiased study of transcription factor binding in the same regions and identified a novel enhancer for the carnitine transporter SLC22A5 (OCTN2) gene. Our results provided insights into the functional relationships between chromatin modifications and regulatory activity in human cells and a new tool for the functional annotation of the human genome.

2.1 Introduction

Activation of eukaryotic gene transcription involves the coordination of a multitude of transcription factors and cofactors on regulatory DNA sequences, such as promoters and enhancers, and the chromatin structure containing these elements[1-3]. Promoters are located at the 5'-ends of genes immediately surrounding the transcriptional start site (TSS) and serve as the point of assembly of the transcriptional machinery and initiation of transcription[4]. Enhancers contribute to the activation of their target genes from positions upstream, downstream, or within a target or neighboring gene[5,6]. Deciphering the regulatory information encoded in the genome will require a thorough understanding of the relationships between the transcriptional activities of these different types of cis-regulatory sequence elements and the epigenetic features of the chromatin surrounding them. Significant progress in the fields of epigenetics and chromatin biology suggests a histone code[7] of ever-increasing complexity with profound implications of chromatin as both receptive substrate and predictive signal in a variety of biological processes[3,8].

Recent investigations using ChIP-chip have described the chromatin architecture of transcriptional promoters in yeast, fly, and mammalian systems[9]. In a manner largely conserved across species, active promoters are marked by acetylation of various residues of histones H3 and H4 (H3ac, H4ac) and methylation of histone H3 lysine 4, particularly tri-methylation of this residue (H3K4me3). Nucleosome depletion is also a general characteristic of active promoters in yeast and flies, though this feature remains to be thoroughly examined in mammalian systems. While some studies suggest that distal regulatory elements like enhancers may be marked by similar histone modification patterns[10-13], the distinguishing chromatin features of promoters and enhancers have yet to be determined, hindering our understanding of a predictive histone code for different classes of regulatory elements. Here, we present high-resolution maps of multiple histone modifications and transcriptional regulators in 30 Mbp of the human genome, revealing that active promoters and enhancers are associated with distinct chromatin signatures that can be used to predict these regulatory elements in the human genome.

2.2 Chromatin architecture and transcription factor localization

ChIP-chip analysis[14] was performed to determine the chromatin architecture along 44 human loci selected by the ENCODE consortium as common targets for genomic analysis[15], totalling 30 Mbp. We investigated the patterns of core histone H3 and five histone modifications: acetylated histone H3 lysine 9/14 (H3ac), acetylated histone H4 lysine 5/8/12/16 (H4ac), and mono-, di-, and tri-methylated histone H3 lysine 4 (H3K4me1, H3K4me2, H3K4me3). We also examined binding of two components of the basal transcriptional machinery (RNAPII and TAF1) and the transcriptional coactivator p300 to identify active promoters and enhancers, respectively. Three biological replicate ChIP-chip experiments were carried out for each marker in HeLa cells before and after treatment with interferon-gamma (IFNγ), as p300 is known to be involved in the cellular response to this cytokine[16]. ChIP samples were amplified, labelled, and hybridized to tiling oligonucleotide microarrays covering the non-repetitive sequences of 30 Mbp at 38-bp resolution. The microarray data were analyzed by standard methods to determine average enrichments for each marker at every probe, generating high-resolution maps of histone modifications and transcriptional regulator binding for 1% of the human genome. To validate our ChIP-chip results, we performed conventional ChIP against RNAPII and tested for enrichment at 121 sites in the ENCODE regions using quantitative real-time PCR, indicating an accuracy of 97%, a specificity of near 100%, and a sensitivity of 82% for our method (see supplementary methods and Table S1). These values are comparable to other ChIP-chip studies[12,17,18] and confirm that our ChIP-chip data is very reliable.

2.3 Chromatin signatures of promoters

To explore chromatin features at human promoters, we examined ChIP-chip profiles along 10 Kbp regions surrounding well-annotated promoters in the ENCODE regions and performed computational clustering to classify each promoter on the basis of histone modification patterns. We examined only those TSSs corresponding to well-annotated RefSeq[19] transcripts for which we had collected expression data, and to prevent interference from neighboring genes, we excluded TSSs within 10 Kbp of each other from the analysis, resulting in a pool of 208 TSSs for clustering; 104 TSSs are defined as active promoters and 104 inactive promoters by gene expression profiling experiments. We observed four distinct classes of promoters in untreated HeLa cells, arranged by the proportion of active promoters within each class (Figure 2.1A, Table S2). Expression levels of transcripts within each class generally increase from class P1 to P4, as most of the inactive promoters are found in class P1 while classes P2-P4 are increasingly composed of active promoters. Average enrichment profiles for each marker within each class (Figure 2.1B) show that occupancy by all five histone modifications, RNAPII, and TAF1 increases at active promoters in a manner related to gene expression levels. Moderate p300 enrichment is also present at many active promoters (a representative active promoter is shown in Figure S1), while the largely inactive class P1 is devoid of any markers. The patterns observed in HeLa cells treated with IFNγ are almost identical (Figure S2A). The transition from H3K4me3 to H3K4me2 to H3K4me1 moving downstream from active promoters into coding regions echoes the pattern seen in small scale studies in human cells[20] and globally in yeast[17,21]. These results confirm previous observations in other organisms that histone modifications are linked to promoter activity.

Interestingly, our analysis revealed a bimodal distribution of all histone modifications centered around peak binding of RNAPII and TAF1 at the TSS, implying depletion of nucleosomes at this position. ChIP-chip data for histone H3 support this conclusion (Figure 2.1A-B, see column H3). Our findings indicate that the nucleosome free region (NFR) observed at promoters in yeast and fly is indeed characteristic of active human promoters, supporting an evolutionarily constrained role for this phenomenon in transcriptional regulation. The degree of nucleosome depletion appears to be related to the level of gene expression, as depletion is not observed in class P1, suggesting that the formation and maintenance of NFRs at active promoters is a regulated process. Distribution around the NFR varies among the histone modifications and promoter classes, but most modifications are found on both sides of the NFR with an asymmetrical bias toward the region immediately downstream, particularly for H3K4me3. H4ac is the only outlier to this trend. The observed histone acetylation and methylation at nucleosomes upstream of the TSS may represent metazoan-specific signatures of chromatin architecture at promoters.

2.4 Chromatin signatures of enhancers

Next, we investigated the chromatin features of human transcriptional enhancers. As previous studies have demonstrated that p300 and related acetyltransferases are present at enhancers (as well as promoters)[10,11], we identified genomic regions in HeLa cells enriched in p300 binding (124 binding sites in untreated cells and 182 sites in treated cells, listed in Table S3) and found that the p300 binding sites exhibit several known features of enhancers. First, the genomic distribution of p300 sites is consistent with the widespread location of enhancers relative to their target genes, as over 75% of p300 binding occurs more than 2.5 Kbp from Gencode[22] known gene 5'-ends (Figure S3A-B). Second, transcriptional regulatory elements such as enhancers have long been known to exhibit increased nuclease sensitivity[23], so we mapped the DNaseI hypersensitive sites (DHSs) in HeLa cells along the ENCODE regions using a recently developed DNase-chip method[24] (Table S4) and found that a significant number of distal p300 sites (69.7%, $p < 1e^{-16}$) overlap with DHSs, representing ~12% of the distal DHSs we identified (Figure S3C). Third, most distal p300 sites are conserved across species; over 60% ($p < 1e^{-16}$) of these sites contain strongly conserved sequence (see supplementary methods). Fourth, a significant number of the distal p300 sites (44.4%, $p = 4.6e^{-15}$) contain independently predicted regulatory modules (PReMods) identified based on clustering of conserved transcription factor binding motifs[25] (see supplementary methods). These lines of evidence provide strong support that the distal p300 binding sites represent a subset of enhancers.

Using the distal p300 binding sites to anchor 10 Kbp regions surrounding each putative enhancer, we performed computational clustering as described above to generate three classes of enhancers (Figure 2.1C-D; classes are arbitrarily named E1-E3 to simplify discussion). Interestingly, we discovered that H3K4me1 is strongly enriched in a broad pattern at nearly all enhancers. This analysis also reveals depletion of histone H3 at enhancers, suggesting that nucleosome depletion is a general feature of both promoters and enhancers, consistent with their DNaseI hypersensitivity. While most active promoters are marked by substantial enrichment of H3K4me3 at the TSS, enhancers generally lack this histone modification. Further, active promoters display a marked depletion of H3K4me1 at the TSS and enrichment of this modification more than 1 Kbp downstream and upstream, while enhancers show strong H3K4me1 enrichment at the peak of p300 binding. H4ac, H3ac, and H3K4me2 are present in varying degrees at both promoters and enhancers, though the bimodal distribution of these modifications observed at active promoters is less pronounced at enhancers. TAF1 and RNAPII are also present at some enhancers, though more weakly than at promoters (reminiscent of the converse weak p300 enrichment at promoters seen in Figure 2.1A-B), suggesting docking of the transcriptional machinery at enhancers or physical interaction between enhancers and active promoters as proposed in various models of enhancer action[5,6]. In spite of some similarities between the histone modification profiles of active promoters and enhancers, the sharp contrasts of their H3K4me1 and H3K4me3 profiles represent distinct chromatin signatures for these different classes of regulatory elements.

2.5 Predicting promoters and enhancers via chromatin signatures

Next, we investigated the possibility that the different chromatin signatures of active promoters and enhancers could be employed to predict these transcriptional regulatory elements in the human genome. Training sets were constructed with histone modification profiles surrounding known TSSs and p300 binding sites in untreated HeLa cells and were used to develop a computational prediction algorithm to locate promoters and enhancers in the ENCODE regions based on similarity to the training set chromatin profiles (Figure 2.2A; see supplementary methods). Our two-stage method of regulatory element identification consists of a primary descriptive prediction followed by secondary discriminative filters (see supplementary methods). To qualify as a high-confidence predicted regulatory element, a region of chromatin must unambiguously match one of the training set profiles.

A total of 198 active promoters (Table S5) were predicted in the ENCODE regions in untreated HeLa cells, clustered as described previously into four classes (named PI-PIV to distinguish them from the known promoters presented in Figure 2.1) (Figure 2.2B). In HeLa cells treated with IFNγ, we predicted 208 promoters (Table S5), with greater than 90% overlap between the untreated and treated prediction sets (Figure 2.2C), supporting the accuracy of our method in identifying promoters in an independent data set. The untreated prediction set contains 140 (79%) of the 177 active RefSeq promoters within the ENCODE regions and 32 (21%) of 155 inactive RefSeq promoters, and 180 predictions (91%) map to known Gencode gene 5'-ends (Figure 2.2D), indicating a high degree of sensitivity and accuracy of promoter prediction. Promoter predictions in treated cells are distributed very similarly (Figure 2.2E). Comparison with the recent RIKEN human CAGE data set[26] reveals that the vast majority of the predicted promoters are supported by multiple CAGE tags (see supplementary methods). Even predicted promoters that do not map to a known Gencode 5'-end are largely supported by multiple CAGE tags (50% in untreated cells, 27% in treated cells) or DHSs (83% in untreated cells, 73% in treated cells). It is possible that the inactive promoters identified in our analysis correspond to transcripts that are expressed at levels below the detection threshold, or these promoters may retain some features of transcriptional competence in the absence of active transcription. Six promoter predictions in untreated HeLa cells (nine predictions in treated cells) do not correspond to any known or putative 5'-ends and likely represent novel promoters; all of these predicted novel promoters overlap with DHSs.

We also predicted 389 enhancers (Table S5) in untreated HeLa cells (Figure 2.3A; enhancer predictions are classified EI-EIV to distinguish them from the p300 binding sites presented in Figure 2.1) and 324 enhancers in treated cells, with 89% overlap between prediction sets (Figure 2.3B). Though the prediction algorithm was trained on the histone modification profiles of untreated cells, it accurately identified 77% of the distal p300 binding sites in IFNγ-treated cells (Figure 2.3E), indicating a high degree of sensitivity for the prediction of enhancers in an independent data set. Several lines of evidence support the function of these predictions as enhancers. First, over 85% of predictions are located more than 2.5 Kbp from known gene 5'-ends (Figure 2.3C, Figure S4), consistent with their predicted function. Second, they are evolutionarily conserved, with 53.3% ($p < 1e^{-16}$) containing a strongly conserved sequence. Third, many overlap with predicted transcriptional regulatory modules (36.3%, $p = 1.7e^{-4}$). Fourth, a significant proportion of the enhancer predictions

(55.3%, $p < 1e^{-16}$) overlap with DHSs, including the well-known HS2 enhancer in the β-globin locus[27] (Figure S5). Of 587 distal DHSs in HeLa cells, we predict that 175 (29.8%) are enhancers; the other distal DHSs likely represent additional regulatory elements such as repressors or insulators, or sequences that contribute to chromatin organization. Finally, 86 enhancer predictions in the untreated set (and 116 in the treated set) map to distal p300 binding sites (Figure 2.3D-E) and many others appear to be enriched in p300 binding, but below the threshold of our target selection.

We also discovered that many predicted enhancers lack p300 binding. To determine if these genomic regions were occupied by p300-independent transcriptional coactivator complexes, we performed additional ChIP-chip experiments to examine binding of TRAP220, a component of the Mediator complex that has been shown to occupy enhancers as well as promoters[10,11]. Of 162 TRAP220 binding sites we identified in the ENCODE regions (Table S6), 78 (48.1%) are located far from known 5'-ends of transcripts and may represent potential enhancers. Almost 63% of the distal TRAP220 sites are contained within our enhancer prediction set (Figure 2.3D), and 18 of them are bound by TRAP220 but not p300, confirming the identity of these predicted enhancers. This result suggests that our chromatin-based prediction model is not limited only to enhancers marked by p300. Overall, the majority of predicted enhancers (63.5%) are supported by DNaseI hypersensitivity, binding of p300, binding of TRAP220, or a combination of these features (Figure 2.3F).

2.6 Identification of a novel enhancer for *SLC22A5*

To confirm the potential of our approach to identify enhancers that regulate the activity of target human promoters, we next examined a predicted enhancer located 6 kbp upstream of the SLC22A5 (OCTN2) gene on chr5 (Figure 2.4A). SLC22A5 is a widely expressed gene that codes for a carnitine transporter [28-31]. Mutations in this gene have been identified as a cause of systematic carnitine deficiency, a condition occurring mostly in children that prevents the body from using fats for energy and can result in symptoms including encephalopathy, cardiomyopathy, hypoglycemia and, in serious complications, heart failure, liver problems, coma, and sudden unexpected death [32-35]. While substantial research has been devoted to the role of SLC22A5 in carnitine transport, fatty acid metabolism and related human diseases, very little is known about the transcriptional regulation of this gene. We cloned a region of the SLC22A5 locus (L) containing the promoter and predicted enhancer (E) into a luciferase reporter construct and compared its activity to that of the locus without the predicted enhancer (LΔE) in transiently transfected HeLa cells. The deletion of the predicted enhancer caused a 2.5-fold reduction in reporter activity (Figure 2.4B), supporting the necessity of this site for full activity of the SLC22A5 promoter. We then cloned the predicted enhancer downstream of the luciferase gene in a construct containing the proximal SLC22A5 promoter (P^S) and observed 4.2-fold greater reporter activity from the promoter-enhancer construct (P^SE) than the construct containing only the promoter (Figure 2.4B), confirming that the predicted enhancer is sufficient to increase the activity of this promoter in a position-independent manner.

The predicted enhancer did not stimulate reporter activity in the absence of the SLC22A5 promoter (data not shown). Our results suggest that the putative SLC22A5 enhancer identified by our method is indeed critical for optimal transcriptional activation of this gene.

2.7 Functional validation of promoter and enhancer predictions

To further assess the accuracy of our predictions, we compared our high-confidence prediction sets to a list of in vivo STAT1 binding sites independently mapped in the ENCODE regions, hypothesizing that STAT1 sites are likely to occupy both promoters and enhancers. We performed ChIP-chip in HeLa cells before and after IFNγ treatment as described above and additionally on a PCR-product microarray platform (see supplementary methods) and validated the results using quantitative real-time PCR, generating a list of 13 high-confidence STAT1 sites in IFNγ-treated cells (Table S7); as expected, no STAT1 binding was detected in cells prior to treatment. We compared the STAT1 sites to our prediction lists and found that seven STAT1 sites map to promoter predictions, four map to enhancer predictions, and two are not near any predictions, indicating that our prediction model is capable of detecting the majority (>84%) of an independently generated collection of putative regulatory elements. Four of the seven promoter predictions map to known TSSs: IRF1 (a known STAT1 target), RPS9, c21orf59, and IFNAR2. All of these genes are expressed in HeLa cells, supporting the accuracy of our active promoter predictions.

To validate the novel promoter and enhancer predictions at STAT1 sites, we examined their functional properties using reporter assays. As two adjacent STAT1 sites on chr5 (STAT1.02, -.03) map to the same promoter prediction, we examined the closer of the two sites along with the other novel STAT1 promoter prediction (Figure 2.5A), four STAT1 enhancer predictions (Figure 2.5B), and the two non-predicted STAT1 sites (Figure 2.5C). To test for promoter activity, regions containing the STAT1 sites were amplified from genomic DNA and cloned upstream of the luciferase gene in vectors lacking a promoter (Figure 2.5D); to test for enhancer activity, the same fragments were cloned downstream of the luciferase gene into vectors containing the SV40 minimal promoter (Figure 2.5E), as enhancers are thought to contribute to target gene activation regardless of their position relative to the gene promoter. Clones were transiently transfected into HeLa cells and assayed for reporter activity before and after treatment with IFNγ.

Both STAT1 promoter predictions stimulated reporter activity in the absence of the SV40 promoter when cloned in the upstream position (Figure 2.5D), in accord with their predicted function. Three STAT1 enhancer predictions (STAT1.08-.10) stimulated strong reporter activity when cloned in the downstream position (Figure 2.5E) but required the presence of the SV40 promoter (see supplemental methods), consistent with the positional-independence and promoter-dependence of enhancer activity. The fourth enhancer prediction (STAT1.11) exhibited only weak enhancer activity, though we noted that the STAT1 site in this region is further away from the prediction (710 bp) than any of the other STAT1 sites that we examined (average

~240 bp). The effect of IFNγ is variable among the different sites in both ChIP-chip binding profiles and reporter activity, though there seems to be a relationship between inducibility of p300 binding and reporter activity. Interestingly, one promoter prediction (STAT1.03) also showed enhancer activity (Figure 2.5E). Examination of our pre-filtering prediction lists (see supplemental methods) revealed a predicted enhancer within the STAT1.03 cloned region, explaining the apparent dual functional activity of this novel promoter. The non-predicted sites (STAT1.12, -.13) displayed no functional activity and were not marked by either of the distinctive histone modification patterns, supporting the specificity of our model. It is still possible that these sites are actually regulatory elements that cannot be tested in our system due to their function or a requirement for native chromatin context, but it is worth noting that these are the only two STAT1 sites that did not exhibit DNaseI hypersensitivity.

These data provide functional validation for our model of distinct chromatin signatures at promoters and enhancers, confirm that our computational approach can accurately predict the position and function of these transcriptional regulatory elements on the basis of their chromatin signatures, and suggest a direct connection between chromatin signatures and the regulatory potential of the DNA sequences that they denote.

2.8 Discussion

In summary, we mapped five histone modifications, four general transcription factors, and nucleosome density at high resolution in 30 Mbp of the human genome, identifying chromatin features that distinguish promoters from enhancers. While both kinds of regulatory elements share some features such as nucleosome depletion and enrichment of histone acetylation and H3K4me2, the high-resolution profiles of these markers and the dichotomy of H3K4me3 and H3K4me1 enrichment at active promoters and enhancers define chromatin signatures that can be used to locate novel regulatory elements in the human genome. The H3K4me1 enhancer signature is present in HeLa cell chromatin at multiple loci whose enhancer activity was functionally validated, including a putative novel enhancer for the SLC22A5 gene.

Previous studies have identified some histone modification patterns of promoters and heterochromatin, but our findings expand the current knowledge of chromatin architecture at human promoters and present evidence for novel chromatin features of human enhancers, representing an effective new strategy for identifying and distinguishing promoters and enhancers. The presence of histone acetylation and methylation that we observe upstream of the TSSs of active human promoters has not been reported in yeast and suggests some transcriptional regulatory mechanisms specific to human gene expression. Additionally, the discovery of H3K4me1 at human enhancers may contribute to our understanding of how enhancers function in tissue-specific gene regulation.

In recent years, the genome sequences of a growing number of organisms have been obtained, but extracting functional information from these nucleotide sequences remains a great challenge, as our knowledge of transcription factor binding motifs is incomplete and current sequence-based computational tools are limited in their ability to predict the regulatory function of genomic sequences. Here, we present a strategy to identify transcriptional regulatory elements on the basis of their epigenetic characteristics, independent of motifs or other sequence features. Our chromatin-based prediction model provides a means to locate and distinguish promoters and enhancers at high-resolution and with high degrees of sensitivity and specificity. Although the prediction model was trained only on data from untreated HeLa cells, the sensitivity of the model in data from IFNγ-treated cells supports the utility of our approach in analyzing independent data sets. The results of the functional assays of predicted STAT1 binding sites confirm the ability of our prediction model to identify the location and function of novel promoters and enhancers, even prior to their activation. Because we employed the histone modification profiles at distal p300 binding sites as the basis for our enhancer prediction strategy, we were initially concerned that our predictions might be biased toward only the subset of enhancers bound by p300. Based on the overlap of our predictions with 63% of distal TRAP220 sites and 30% of distal DHSs, however, we conclude that our model is not biased towards p300 binding sites and that the chromatin signatures we observed are not limited to this subset of enhancers. Extension of our model to additional cell types and other components of chromatin architecture will be useful in determining the mechanisms of enhancer maintenance and function in regulating tissue-specific gene expression, findings which will be particularly important to our knowledge of how epigenetic factors and distal transcriptional regulatory elements contribute to human development and disease.

Our approach will also be valuable to the functional annotation of the human genome, as it provides a novel and effective means to locate active transcriptional enhancers that have thus far eluded identification on a large scale. Given the degree of structural and functional conservation of chromatin and histone modifications from yeast to humans, these predictive chromatin signatures may be useful in annotating promoters and enhancers in the genomes of a variety of organisms.

2.9 Methods

For detailed methods and materials, please refer to Supplementary Information

Briefly, HeLa cells were cultured under adherent conditions in DMEM + 10% FBS. Three biological replicates of IFNγ-treated and untreated cells were crosslinked and harvested as previously described[18], except that cells were crosslinked for 20 minutes at 37ºC. Chromatin preparation, ChIP-chip, DNA purification, and LM-PCR were performed as previously described[18] using commercially available antibodies (α-histone H3, Abcam ab1791; α-H4ac, Upstate 06-866; α-H3ac, Upstate 06-599; α-H3K4me1, Abcam ab8895; α-H3K4me2, Upstate 07-030; α-H3K4me3, Upstate 07-473; α-RNAPII, Covance MMS-126R; α-TAF1, Santa Cruz sc-735; α-p300, Santa Cruz sc-585; α-TRAP220, Santa Cruz sc-5334). ChIP-DNA samples were labelled and hybridized to NimbleGen ENCODE HG17 microarrays (NimbleGen Systems, Inc.). Data were analyzed using standard methods, and ChIP-chip targets for RNAPII were selected with the Mpeak program and validated by quantitative real-time PCR using the iCycler™ and SYBR-green iQ™ Supermix (Bio-Rad Laboratories). Gene expression in treated and untreated HeLa cells was analyzed using HGU133 Plus 2.0 microarrays (Affymetrix) as described[18]. DNase-chip was performed and the data analyzed as described[24]. Promoters (TSSs) and putative enhancers (p300 binding sites) were clustered on the basis of histone modification patterns using K-means clustering of 10 Kbp windows centered on each target. Average profiles were generated for each class of promoter and enhancer and used to train a computational prediction model to identify promoters and enhancers on the basis of histone modification ChIP-chip profiles. Predictions were further filtered by correlation to chromatin signatures to remove false positives and ambiguous classifications. Predicted regulatory modules were obtained from http://genomequebec.mcgill.ca/PReMod/, phastCons and CAGE data were extracted from the UCSC Genome Browser, and binding sites and predictions were mapped relative to the October 2005 hg17 Gencode gene sets. ChIP-chip was performed against STAT1 (α-STAT1, Santa Cruz sc-345) as described above and using PCR microarrays, and the results were validated by qRTPCR. Predicted STAT1 sites were cloned into modified pGL3 reporter constructs (Promega), transiently transfected into HeLa cells, and assayed for luciferase activity before and after IFNγ treatment using the Dual Luciferase Kit (Promega). Raw and processed data for the microarray experiments can be found at GEO (Accession number GSE6273), the UCSC genome browser (http://genome.ucsc.edu), and http://licr-renlab.ucsd.edu/download.html. All supplementary tables (S1 – S10) are available at http://licr-renlab.ucsd.edu/download.html.

2.10 Acknowledgment

Chapter 2, in full, is a reprint of the material as it appears in “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.” Nature Genetics 2007, Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, and Ren B. The dissertation author was the primary investigator and author of this publication.

We thank Dr. Jim Kadonaga, Dr. Xiangdong Fu, and members of the Ren lab for comments. This work was supported by funding from LICR (BR), NHGRI (BR, ZW, and RDG), and NCI (BR).

NDH, BR, and RDG designed the transcription factor and histone ChIP-chip experiments; GEC designed and performed the DNase-chip experiments; NDH, RKS, CWC, RDH and SVC conducted the ChIP-chip experiments; NDH, GH, LOB, KAC, and CQ analyzed the microarray data; GH, NDH, BR and WW conceived and developed the promoter and enhancer prediction method. Independently, YF and ZW discovered the promoter-associated chromatin signatures. NDH and BR wrote the manuscript.

2.11 Figures

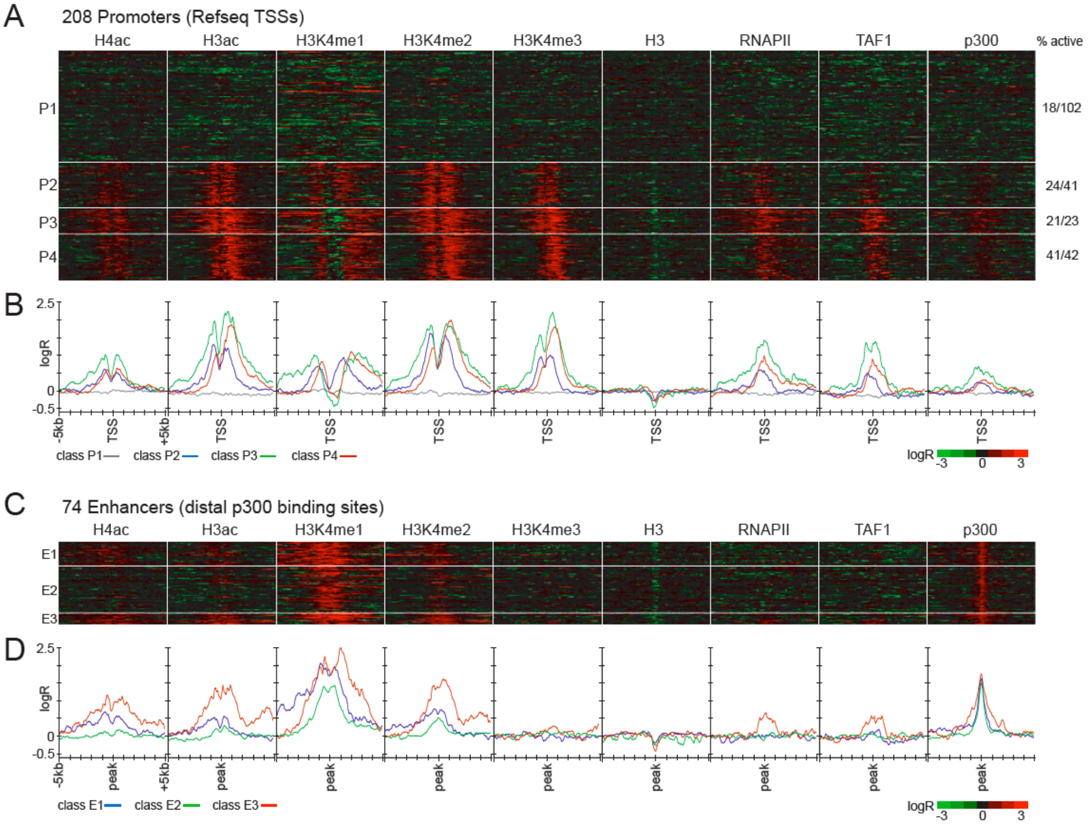

Figure 2.1. Features of human transcriptional promoters and enhancers
ChIP-chip was performed against six histone markers and three general transcription factors in the ENCODE regions and the data were clustered to reveal patterns at annotated promoters (A) and distal p300 binding sites (C). Promoter clustering was performed with 10 Kbp windows centered on RefSeq TSS; enhancer windows were centered on promoter-distal p300 binding peaks. Average profiles for each marker within each class are shown below the clusters (B, D); each class is represented by a different color. The proportion of expressed genes (% active) in each promoter class is presented to the right of the cluster, illustrating the relationship between histone modification patterns and gene expression. Comparison of the clusters shows that active promoters and enhancers are similarly marked by nucleosome depletion (see column H3) but distinctly marked by mono- and tri-methylation of histone H3 lysine 4 (see columns H3K4me1 and H3K4me3). Note the depletion of H3K4me1 and the peak of H3K4me3 at the TSS in promoters, compared to the enrichment of H3K4me1 and lack of H3K4me3 at enhancers. The presence of histone methylation and acetylation upstream and downstream of the TSS at promoters is distinct from the primarily downstream localization of these markers observed at yeast promoters. The same procedure was performed on data from treated HeLa cells, yielding similar results (Figure S2).

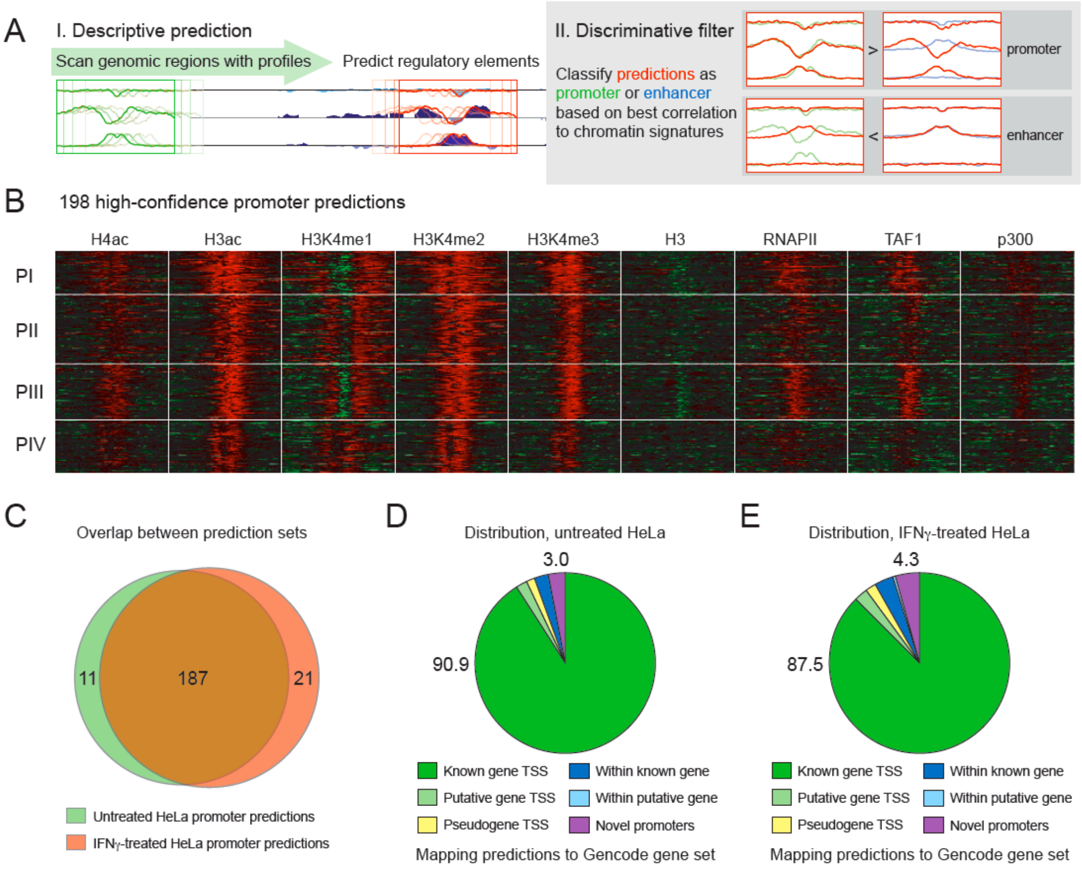

Figure 2.2. Prediction of promoters based on chromatin signatures
(A) A general scheme of the prediction method. I: Features of established transcriptional promoters (and enhancers) were analyzed to yield descriptive histone modification profiles used in scanning genomic regions for novel regulatory elements. II: Predictions were filtered and classified as promoters or enhancers based on correlation with H3K4me1 and H3K4me3 chromatin signatures. (B) 198 active promoters were predicted in the ENCODE regions in untreated HeLa cells and clustered into classes PI-PIV. The predictions contain 140 active RefSeq promoters and 32 inactive RefSeq promoters, indicating a sensitivity of 79.1% for active promoter detection. (C) The high degree of overlap between untreated and IFNγ-treated HeLa promoter prediction sets supports the applicability of our approach to independent data sets. The majority of predicted promoters map to known Gencode 5'-ends in untreated (D) and treated cells (E), confirming the accuracy of our predictions.

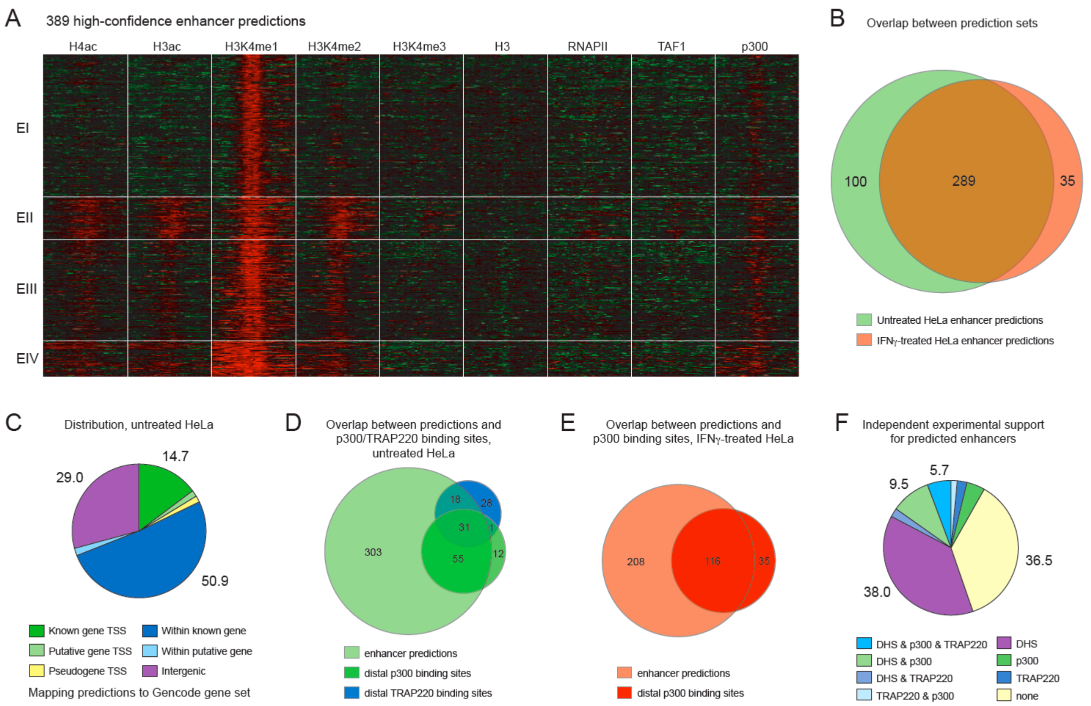

Figure 2.3. Prediction of enhancers based on chromatin signatures
(A) 389 enhancers were predicted in the ENCODE regions in untreated HeLa cells and clustered into classes EI-EIV. The predicted enhancers display the H3K4me1 enrichment and lack of H3K4me3 observed at distal p300 binding sites. (B) The high degree of overlap between untreated and IFNγ-treated HeLa enhancer prediction sets supports the applicability of our approach to independent data sets. The majority of enhancer predictions in untreated (C) and treated cells (see Figure S4) are found away from known Gencode 5'-ends, similar to p300 binding site distribution. The enhancer prediction sets contain the majority of known distal p300 binding sites in untreated (D) and treated cells (E), confirming the sensitivity of our approach even though the prediction algorithm was trained only on data from untreated cells. (F) Most enhancers predicted on the basis of their chromatin signatures are also supported by DNaseI hypersensitivity (DHS) and/or binding of p300 and/or TRAP220.

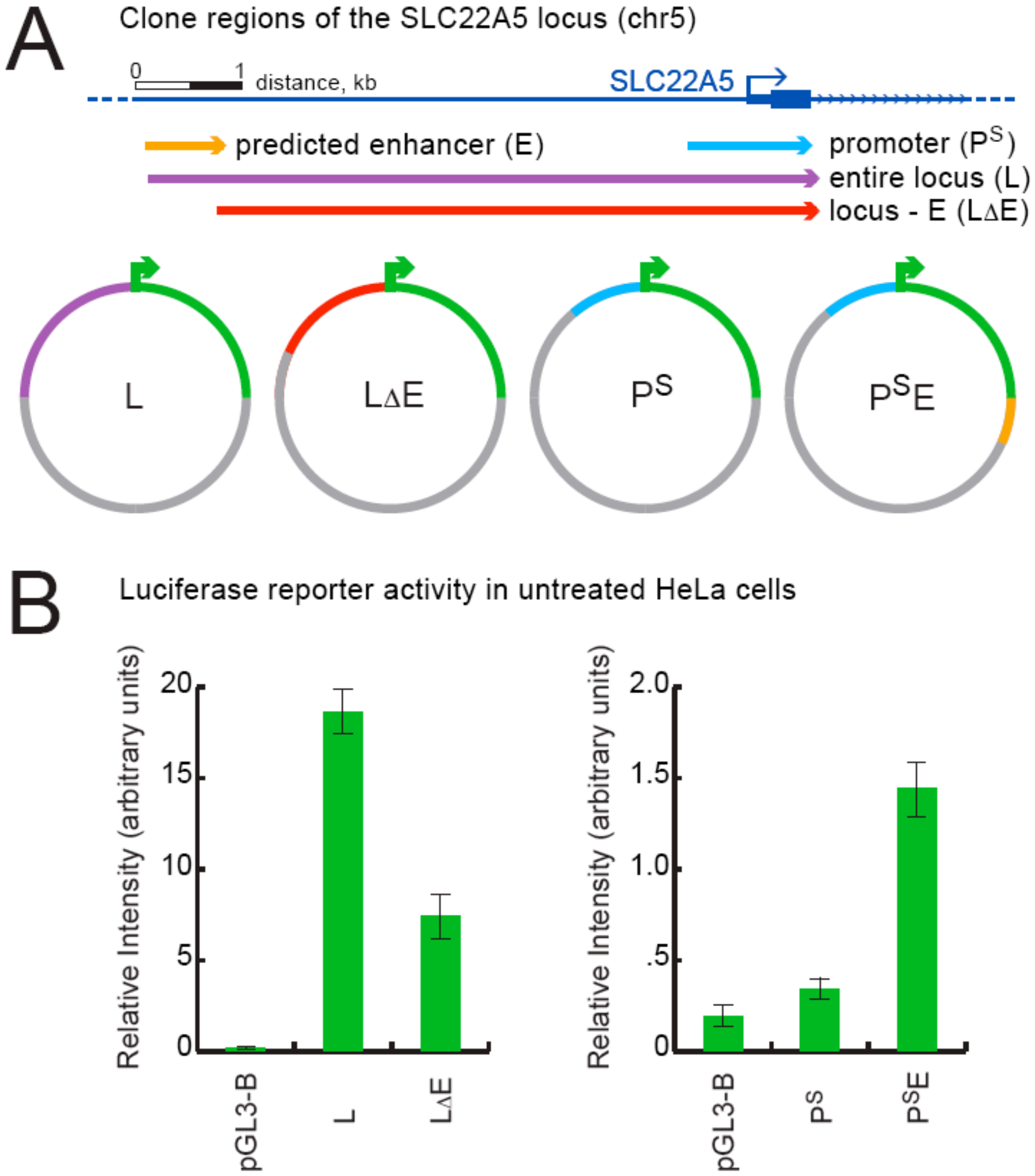

Figure 2.4. Identification of a putative novel enhancer for the SLC22A5 gene
(A) To test the effect of a predicted enhancer on a nearby promoter, regions of the SLC22A5 locus were cloned into pGL3 reporter constructs in the direction indicated. (B) Luciferase activity of the entire 6.5 kbp locus (L) was reduced 2.5-fold by the deletion of 700 bp containing the predicted enhancer (LΔE), and the presence of the predicted enhancer downstream of the luciferase gene in a construct containing the SLC22A5 promoter (P^SE) caused a 4-fold increase in activity compared to the promoter alone (P^S) (error bars represent s.d.).

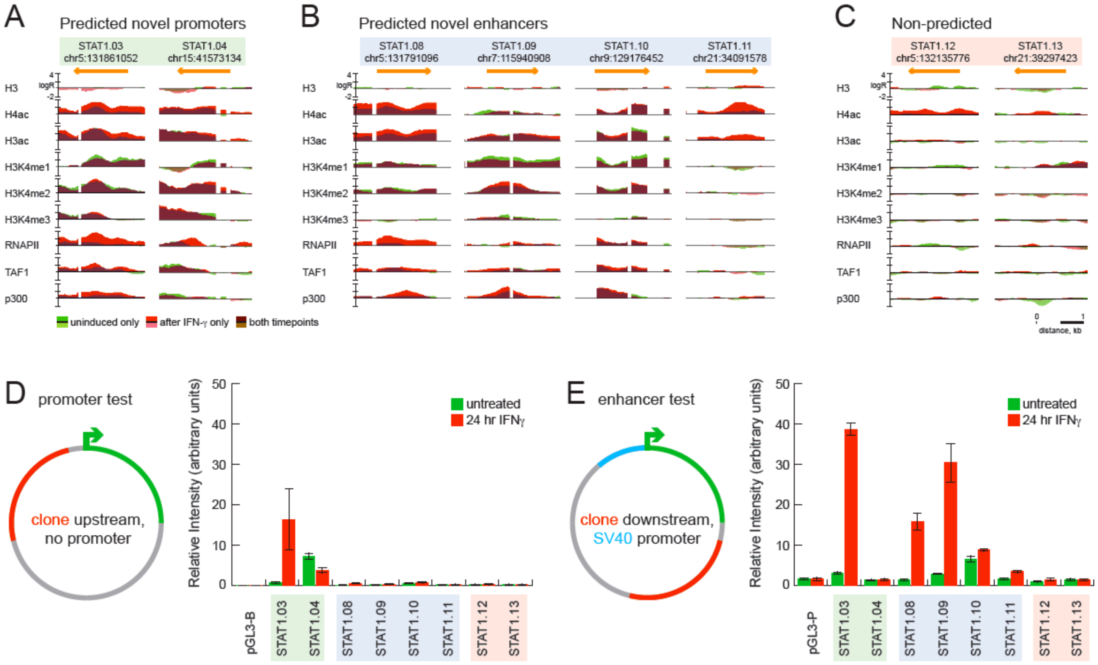

Figure 2.5. Validation of the prediction model by STAT1 binding and reporter assays

Of 13 high-confidence STAT1 binding sites, four are found at known promoters (not shown), two at predicted novel promoters (A), four at predicted enhancers (B), and two are not predicted (C). The eight STAT1 sites in A-C were cloned into pGL3 reporter constructs to examine their regulatory potential as promoters (D) and enhancers (E) (error bars represent s.d.). The coverage and direction of each clone are represented by orange arrows in A-C, and ChIP-chip profiles of each marker are shown at the eight STAT1 binding sites, before and after IFNγ treatment (green and red, respectively, where brown indicates enrichment at both time points; images generated in part at http://genome.ucsc.edu using hg17). The STAT1 binding sites in (A) and (B) function as predicted in the reporter assays, while the non-predicted sites in (C) display no reporter activity.

2.12 References

1. Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 2000;14:2551-69.

2. Orphanides G, Reinberg D. A unified theory of gene expression. *Cell* 2002;108:439-51.

3. Nightingale KP, O'Neill L P, Turner BM. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr Opin Genet Dev* 2006;16:125-36.

4. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. *Annu Rev Biochem* 2003;72:449-79.

5. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. *Science* 1998;281:60-3.

6. Bulger M, Groudine M. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev* 1999;13:2465-77.

7. Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000;403:41-5.

8. Margueron R, Trojer P, Reinberg D. The key to development: interpreting the histone code? *Curr Opin Genet Dev* 2005;15:163-76.

9. Barrera LO, Ren B. The transcriptional regulatory code of eukaryotic cells - Insights from Genome-wide Analysis of Chromatin Organization and Transcription Factor Binding. *Curr Opin Cell Bio* 2006;In press.

10. Hatzis P, Talianidis I. Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol Cell* 2002;10:1467-77.

11. Wang Q, Carroll JS, Brown M. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell* 2005;19:631-42.

12. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, 3rd, Gingeras TR, Schreiber SL, Lander ES. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* 2005;120:169-81.

13. Roh TY, Cuddapah S, Zhao K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* 2005;19:542-52.

14. Kim TH, Ren B. Genome-wide Analysis of Protein-DNA Interactions. *Annual Review of Genomics and Human Genetics* 2006;7.

15. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306:636-40.

16. Horvai AE, Xu L, Korzus E, Brard G, Kalafus D, Mullen TM, Rose DW, Rosenfeld MG, Glass CK. Nuclear integration of JAK/STAT and Ras/AP-1 signaling by CBP and p300. *Proc Natl Acad Sci U S A* 1997;94:1074-9.

17. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, Zeitlinger J, Lewitter F, Gifford DK, Young RA. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 2005;122:517-27.

18. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. *Nature* 2005;436:876-80.

19. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005;33:D501-4.

20. Kouskouti A, Talianidis I. Histone modifications defining active genes persist after transcriptional and mitotic inactivation. *Embo J* 2005;24:347-57.

21. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ. Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* 2005;3:e328.

22. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 2006;7 Suppl 1:S4 1-9.

23. Felsenfeld G. Chromatin unfolds. *Cell* 1996;86:13-9.

24. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* 2006;3:503-9.

25. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* 2006;16:656-68.

26. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;38:626-635.

27. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. Locus control regions. *Blood* 2002;100:3077-86.

28. Schomig E, Spitzenberger F, Engelhardt M, Martel F, Ording N, Grundemann D. Molecular cloning and characterization of two novel transport proteins from rat kidney. *FEBS Lett* 1998;425:79-86.

29. Sekine T, Kusuhara H, Utsunomiya-Tate N, Tsuda M, Sugiyama Y, Kanai Y, Endou H. Molecular cloning and characterization of high-affinity carnitine transporter from rat intestine. *Biochem Biophys Res Commun* 1998;251:586-91.

30. Tamai I, Ohashi R, Nezu J, Yabuuchi H, Oku A, Shimane M, Sai Y, Tsuji A. Molecular and functional identification of sodium ion-dependent, high affinity human carnitine transporter OCTN2. *J Biol Chem* 1998;273:20378-82.

31. Wu X, Prasad PD, Leibach FH, Ganapathy V. cDNA sequence, transport function, and genomic organization of human OCTN2, a new member of the organic cation transporter family. *Biochem Biophys Res Commun* 1998;246:589-95.

32. Nezu J, Tamai I, Oku A, Ohashi R, Yabuuchi H, Hashimoto N, Nikaido H, Sai Y, Koizumi A, Shoji Y, Takada G, Matsuishi T, Yoshino M, Kato H, Ohura T, Tsujimoto G, Hayakawa J, Shimane M, Tsuji A. Primary systemic carnitine deficiency is caused by mutations in a gene encoding sodium ion-dependent carnitine transporter. *Nat Genet* 1999;21:91-4.

33. Shoji Y, Koizumi A, Kayo T, Ohata T, Takahashi T, Harada K, Takada G. Evidence for linkage of human primary systemic carnitine deficiency with D5S436: a novel gene locus on chromosome 5q. *Am J Hum Genet* 1998;63:101-8.

34. Stanley CA. Carnitine deficiency disorders in children. *Ann N Y Acad Sci* 2004;1033:42-51.

35. Wang Y, Ye J, Ganapathy V, Longo N. Mutations in the organic cation/carnitine transporter OCTN2 in primary carnitine deficiency. *Proc Natl Acad Sci U S A* 1999;96:2356-60.

Chapter 3

A Genome-wide Map of Human Transcriptional Enhancers

This chapter includes supplementary figures, supplementary tables, and supplementary information and data files, all of which are available at http://bioinformatics-renlab.ucsd.edu/enhancer

Abstract

Transcriptional enhancers play critical roles in cell type-specific regulation of gene expression, but our knowledge of these *cis*-regulatory elements in the human genome is still incomplete. We previously developed a method to identify transcriptional enhancers based on their unique chromatin signatures. Here, we describe the first genome-wide map of enhancers obtained using this method in human cells. The enhancers we identified are strongly correlated to cell type-specific gene expression patterns on a global scale. We observed significant enrichment of diverse

transcription factor binding motifs in enhancers and also identified 22 novel enhancer-specific DNA sequence motifs. In addition, we found that many enhancers are marked by characteristic histone modifications even prior to binding of sequence-specific activators that exert regulatory effects. Remarkably, this genome-wide map of enhancers also correctly predicts a significant portion of the *in vivo* binding sites for diverse activators in distinct cell types and conditions, suggesting an epigenetic mechanism to retain specific chromatin modifications at some enhancers during lineage specification. Our results significantly expand the current catalog of human enhancers and provide new insights into the global properties of enhancers and their role in cell type-specific gene expression.

3.1 Introduction

Transcriptional regulation of eukaryotic gene expression is a complex process that requires precise spatial and temporal coordination of a host of regulatory inputs, including DNA sequence elements, transcription factor and coactivator binding, and chromatin structural features, all of which cooperate to activate transcription from promoter sequences located at the 5' end of each gene[1-4]. Complicating our understanding of this process, however, is the difficulty of locating the many additional *cis*-regulatory DNA sequence elements responsible for appropriate modification and maintenance of gene expression patterns, including enhancers, silencers, locus control regions, and insulators[2]. A complete catalog of these *cis*-

regulatory elements is necessary to develop a better mechanistic understanding of transcriptional regulation.

Various strategies have been employed to locate transcriptional regulatory elements on a genome-wide scale. Comparative genomics techniques search for DNA sequences that are similar or identical across evolutionarily distinct species, on the hypothesis that such conserved sequences are maintained under selective pressure to perform regulatory functions[5,6]. Such efforts have located multiple potential regulatory elements, many of which have been experimentally verified *in vivo*[7,8]. Other recently developed high-throughput methods interrogate the structure of chromatin for regions that are sensitive to DNaseI[9,10], as these DNaseI hypersensitive sites (DHS) have long been known to mark active *cis*-regulatory elements[11,12]. This strategy has yielded DHS maps of several regions of the human genome in a variety of cell types[9,10,13]. Other experimental methods including ChIP-chip, GMAT, and ChIP-PET[14], as well as the recently developed ChIP-Seq[15,16] have identified regions of DNA-protein interaction on a large scale, providing maps of sequence-specific transcription factor binding sites and histone modifications that contribute to regulation of gene expression[17-19].

While effective at locating regions with transcriptional regulatory potential, these techniques share a deficiency in distinguishing different classes of regulatory elements. For example, DHS are found at promoters and enhancers alike, as are transcription factor binding sites and hyperacetylated histones, and the degeneracy of many DNA sequence elements prevents their classification into different regulatory categories. Furthermore, an increasing number of completed genome sequences and comparative genomics studies suggest that many DNA elements that are not conserved from human to lower organisms are responsible for the elegant transcriptional regulatory processes that contribute most to making us human[20,21].

We previously reported that transcriptional enhancers throughout 1% of the human genome (the ENCODE regions[22]) were strongly enriched in mono-methylation of histone H3 lysine 4 (H3K4me1), while core promoters lack this particular histone modification and are instead marked by tri-methylation of the same residue (H3K4me3)[13]. These differences form the basis of distinct chromatin signatures that we employed to unambiguously predict hundreds of active promoters and enhancers in the ENCODE regions, many of which we functionally validated using reporter assays. In this study, we extend our prediction strategy to the entire human genome, generating the first genome-wide map of transcriptional enhancers based on these chromatin signatures. This map reveals global properties of enhancers and their role in cell type-specific gene expression. Intriguingly, we find that a large number of enhancers are already poised for activation prior to binding of specific activator proteins, providing evidence for epigenetic mechanisms to retain histone modification marks during cellular proliferation and differentiation. The genome-wide enhancer map may be viewed in the UCSC Genome Browser via http://bioinformatics-renlab.ucsd.edu/enhancer

3.2 Genome-wide prediction of active promoters and enhancers based on chromatin signatures

We performed ChIP-chip in HeLa cells to map enrichment patterns throughout the entire human genome as previously described[23,24] for mono- and tri-methylation of histone H3 lysine 4 (H3K4me1 and H3K4me3, respectively), identifying 13116 active promoters (Figure S1A, Table S1) and 38716 enhancers (see below) in the HeLa genome using our chromatin signature-based prediction algorithm (see Methods)[13]. We found that 9835 (75%) predicted promoters overlap with 5' ends of UCSC Known Genes[25] (Figure S1B). We also compared the promoter predictions to the RIKEN human CAGE data set[26] and observed that 11001 (83.9%) overlap with multiple CAGE tags (see Methods). Further, our prediction model correctly located 76% of active RefSeq transcription start sites (TSS) (Figure S1C) and even 31.5% of inactive TSS (see Methods). We also examined the overlap of predicted promoters with CpG islands (as annotated at the UCSC genome browser[27]), sequence elements conventionally understood to be associated with many promoters[1]. The vast majority of active promoter predictions (11186, 85.1%) overlap CpG islands, representing almost half (43.3%) of the genome's CpG islands within the annotated list (Figure S1D). These findings agree with our previous genome-wide promoter analysis[28] and are comparable to the specificity and sensitivity of our prediction model in the ENCODE regions[13], indicating that our genome-wide promoter map is very reliable.

The genome-wide enhancer map is also highly accurate. We identified several previously characterized enhancers based on their chromatin signatures, including the β-globin HS2 enhancer[29], a distal downstream enhancer for the PAX6 gene[30], and a distal upstream enhancer for the PLAT (t-PA) gene[31] (Figure 3.1A). We designed condensed enhancer microarrays to confirm the H3K4me1 and H3K4me3 signatures of the predicted enhancers (see Methods), verifying 36589 (94.5%) enhancers (Figure 3.1B, Table S2) (see Methods). The signature-verified enhancers are distinct from promoters: only 1071 (2.9%) enhancers overlap with Known Gene 5' ends (Figure 3.1C), 3271 (8.9%) overlap with multiple CAGE tags, and 933 (2.5%) overlap with CpG islands. Indeed, the vast majority of predicted enhancers are distal to promoters, with predominantly intronic (37.9%) or intergenic (56.3%) localization (Figure 3.1C). Further, the predicted enhancers are distinct from other distal regulatory elements. Comparison to a genome-wide binding profile of the repressor NRSF/REST[16], which would be expected to bind transcriptional silencer elements, revealed that less than 3% of distal NRSF/REST binding sites overlap predicted enhancers in our map (data not shown), significantly lower than would be expected by random (3.5-fold depletion, P = $1.66e^{-18}$). These findings confirm that our chromatin signature-based prediction method accurately and specifically identifies enhancers.

3.3 Distinct classes of enhancers

To further characterize the enhancer predictions, we used the condensed enhancer microarray to examine DNaseI hypersensitivity (DHS, via DNase-chip), acetylation of H3K27 (H3K27ac, via ChIP-chip), and binding of the transcriptional coactivator p300 and the Mediator component TRAP220 (via ChIP-chip), as these features have been shown to localize at enhancers[13,32,33][Hon G, Hawkins RD et al., manuscript in preparation]. These data were used to cluster the enhancers into four distinct classes, referred to as EI, EII, EIII, and EIV to simplify discussion (Figure 3.1B) (see Methods). Most predicted enhancers (61.4%, consisting of classes EI-EIII) are marked by moderate or high levels of acetylation of H3K27. We found significant p300 and TRAP220 binding at 10741 (29.4%) and 5764 (15.8%) enhancer predictions, respectively, mainly in classes EI and EII but with weaker binding to some other predicted enhancers (see Methods). Additionally, 19776 (54.1%) of the predicted enhancers exhibit significant DNaseI hypersensitivity (see Methods). Collectively, we found that 23722 (64.8%) predicted enhancers are supported by some combination of DHS and/or binding of p300 and/or TRAP220 (Figure 3.1D). The enrichment of these features appears to correlate with H3K27ac levels, as the most hyperacetylated enhancers in classes EI and EII also show the strongest DHS and binding of p300 and TRAP220, while class EIV mostly lacks these marks, and class EIII is intermediate.

Consistent with our previous findings[13] and with the recent ENCODE study[21], we observed that most of our predicted enhancers show little or no H3K4me3 (Figure 3.1B), including the known enhancers that we recovered (Figure 1A). The weak H3K4me3 enrichment seen at some enhancers in classes EI-EIII is possibly a reflection of their activity and physical proximity to target promoters via chromosome looping[2]. This agrees with our previous observation that predicted enhancers with weak H3K4me3 enrichment are also weakly marked by components of basal transcriptional machinery[13].

3.4 Enhancers are globally related to specificity of gene expression

Enhancers are thought to contribute to cell type-specific expression of genes[34]. To examine the relationship of the predicted enhancers to HeLa cell-specific gene expression, we first ranked genes by the specificity of their expression levels in HeLa as compared to expression profiles in three other cell lines (K562, GM06990, and IMR90 cells, representing leukemia, lymphoblast, and fibroblast lineages, respectively), defining the specificity of expression using a function of Shannon entropy as previously described[35] (see Methods). Next, we divided the genome into insulator-defined domains based on published CTCF binding sites in IMR90 cells[23]. We then examined the localization of predicted enhancers in each domain relative to promoters of genes with expression patterns of varying cell type-specificity (see Methods).

We observed a striking enrichment of enhancers in the domains of HeLa-specific expressed genes relative to non-specific expressed genes, HeLa-specific unexpressed genes, and a random distribution (Figure 3.2A), supporting the role of these predicted enhancers in regulating cell type-specific gene expression. Noting that most enhancer enrichment occurred within 200 kb of gene promoters, we counted enhancers within this window (and also within the same insulator-defined domain) around each promoter and compared enhancer counts around HeLa-specific expressed genes, non-specific and unexpressed genes, and a random distribution (see Methods). We observed a 1.7-fold enrichment ($P = 1.19e^{-304}$) of enhancers around HeLa-specific expressed genes relative to random, while enhancers are actually depleted around non-specific and unexpressed genes, 1.2-fold (P = 1.06e^{-9}) and 1.1-fold (P = 4.88e^{-4}), respectively (Figure 3.2C).

We conclude that the predicted enhancers are significantly enriched near expressed genes but not unexpressed genes, and that the general localization of enhancers in HeLa cells corresponds to genes whose expression patterns are specific to HeLa as compared to three other cell types. These findings confirm, on a genome-wide scale, the concept that enhancers regulate cell type-specific gene expression patterns.

3.5 Different classes of enhancers show distinct relationships to gene expression

The variable patterns of H3K27ac, p300, TRAP220, and DHS among the four enhancer classes prompted us to examine the relationship of each class to HeLa-specific gene expression. The high levels of H3K27ac, DHS, p300 and TRAP220 were very similar between classes EI and EII (the differences mainly resting in their asymmetry), so we combined EI and EII for this analysis. We observed that the different classes of enhancers were related to HeLa-specific gene expression patterns in distinct ways. Class EI/II enhancers were very strongly enriched around HeLa-specific expressed genes (Figure 3.2B); when we counted EI/II enhancers around promoters as above, we observed a striking 2.5-fold enrichment (P = 2.97e^{-293}) of enhancers around HeLa-specific expressed genes relative to random (Figure 3.2C). Class EIV enhancers were not as strongly related to HeLa gene expression patterns, with 1.2-fold enrichment ($P = 1.04e^{-10}$) around HeLa-specific expressed genes (Figure 3.2B-C). Class EIII enhancers showed enrichment levels intermediate to those observed for EI/II and EIV (Figure 3.2B), with 1.6-fold enrichment ($P = 4.81e^{-81}$) around HeLa-specific expressed genes (Figure 3.2C). Enhancers in all three classes were depleted around non-specific and unexpressed genes (Figure 3.2C).

The MET proto-oncogene has been implicated in a variety of carcinomas (including cervical)[36,37] and is highly expressed in HeLa cells (a cervical adenocarcinoma) relative to the other cell types studied (at least 84-fold higher expression in HeLa). We compared the chromatin signatures in the MET locus in HeLa cells with those seen in K562 and GM06990 cells [Hon G, Hawkins RD et al., in preparation] and found that 11 HeLa-specific enhancers are predicted near MET (Figure 3.2D), while no enhancers are predicted there in K562 or GM06990 cells. An adjacent gene, CAPZA2, is expressed at more similar levels in all three cell types and is not marked by HeLa-specific enhancers. The specificity of the predicted enhancers depicted in this locus is further supported by H3K27ac and DHS data for these cell types[9,13][Crawford GE, in preparation] (Figure 3.2D).

In addition to the H3K4me1 chromatin signature, enhancers that are most frequently located near HeLa-specific expressed genes (mainly classes EI/II and EIII) exhibited greater enrichment of H3K27ac, DHS, p300 and TRAP220, while enhancers with lower levels of these markers (mainly class EIV) are less strongly associated with these genes. Perhaps some of these enhancers are “poised” rather than active; that is, they are awaiting additional regulatory input (such as binding of a sequence-specific transcription factor) required for their activation, but are already marked by the H3K4me1 enhancer signature (see following sections and Discussion). Interestingly, two of the known enhancers identified by the prediction model (β-globin HS2 and PLAT, Figure 3.1A) fall into class EIV in HeLa cells; a third (PAX6) is in class EIII.

3.6 Identification of novel sequence motifs in enhancers

We examined the predicted enhancers for the presence of DNA sequence elements that may guide the establishment and maintenance of chromatin structure or the recruitment of regulatory factors. We reasoned that if such functional motifs are abundant within enhancers, they could show increased evolutionary conservation across related mammals. Indeed, we found that enhancers show strong 'motif-like conservation', evaluated as the fraction of randomly-sampled motif instances which are conserved, allowing for limited motif loss, incorrect alignments and sequencing errors (see Methods). Enhancers show 4.3% motif-like conservation, which is substantially greater than for remaining intergenic regions (2.9%, $P < 1e^{-100}$) and even promoter regions (3.9%). By contrast, a simple measure of nucleotide identity does not show a significant increase for either promoters or enhancers. This suggests that evolutionary selection in enhancer regions is specifically acting at the motif level, which may help to explain the low sequence conservation for many functional elements identified by the ENCODE project[21].

Consequently, we asked whether motifs for known transcriptional regulators show increased abundance and conservation in enhancer regions. We tested a list of

123 unique TRANSFAC motifs as reported previously[38] (see Methods) and found that 67 (54%) of these motifs are over-conserved in enhancers, and 39 (32%) are enriched in enhancers (Table S3). This suggests that many known motifs for well-studied transcriptional regulators at promoters are likely to also play roles in enhancers, implying strongly shared regulatory mechanisms between these two classes of elements at the DNA sequence level. Indeed, of the 67 known motifs over-conserved in enhancers and the 65 over-conserved in promoters, 54 are over-conserved in both (83%). The enriched motifs include known sequence motifs for binding of transcription factors involved in diverse cellular processes.

Additionally, we searched for evidence of unique enhancer-specific sequence motifs that have previously remained elusive due to the lack of genome-wide knowledge of enhancers. We performed *de novo* motif discovery in enhancer regions using multiple alignments of ten mammalian genomes (see Methods), revealing 41 enhancer motifs, of which 19 match known transcription factor motifs while 22 are novel (Table 3.1). These motifs show conservation rates between 7% and 22% in enhancers (median 9.3%), compared to only 1.1% for control motifs of identical composition. Even without taking conservation into account, 27 (65%) of these motifs show significant enrichment in human enhancers. Further, over 90% of these motifs appear to be unique to enhancers, as only 4 motifs are enriched in promoter regions and 12 are in fact depleted in promoters (Table 3.1). In contrast, shuffled versions of these motifs show significantly reduced enrichment in enhancer regions (only 12% of shuffled motifs, a 5-fold reduction) and also reduced depletion in promoters (22%, a 2-

fold increase). This indicates that although enhancer regions contain many known promoter motifs, they also contain unique regulatory sequences that may be specific to enhancer function.

3.7 A subset of poised enhancers

We were intrigued to find that many of the sequence motifs present in enhancer regions correspond to transcription factors not known to be specific to HeLa cells (for example MYOD, NF-AT, ER, and STAT1). In light of our discovery that a subset of enhancers does not appear to correlate as strongly with cell type-specific expression of genes, this observation suggests that some predictions correspond to enhancers that may be active in other cell types or conditions. To test this hypothesis, we compared the enhancer predictions with the results of several genome-wide studies of binding sites for sequence-specific transcription factors in different cell types.

One investigation reported 3665 estrogen receptor (ER) binding sites in MCF7 cells[39], 3381 (92.2%) of which are located distal to Known Gene 5' ends. Although our enhancer prediction map was generated based on chromatin signature data in HeLa cells, we found that 1173 (34.3%, $P < 1e^{-200}$) distal ER binding sites in MCF7 cells overlap predicted enhancers (Figure 3.3A); two of the ER binding sites overlapping predicted enhancers (both in class EIV) were previously demonstrated to exhibit functional activity in reporter assays[40] (Figure 3.3B). We observed similar overlap for two additional transcription factors, predicting enhancers at 24.2% ($P = 3.68e^{-27}$) of distal p53 binding sites in HCT116 cells[41] (Figure 3.3C) and 28.8% (P <

$1e^{-200}$) of distal p63 binding sites in ME180 cells[42] (Figure 3.3D). In light of the considerable evidence for tissue-specificity of enhancers, it is remarkable that we predict such significant fractions of potential enhancers bound by transcription factors in other cell types. This is in sharp contrast to the significant depletion of the repressor NRSF/REST at the predicted enhancers, as noted earlier.

The results above strongly indicate that the enhancer prediction map includes many enhancers that may be active in other lineages or cellular contexts and yet are marked by enhancer chromatin signatures in HeLa cells. Hypothesizing that these enhancers are poised in HeLa cells and awaiting activation in other cell types or under different physiological conditions, we treated HeLa cells with the cytokine interferon-gamma (IFNγ) and identified STAT1 binding sites throughout the genome using ChIP-chip (see Methods). As a signal-dependent, latent cytoplasmic transcription factor, STAT1 is generally understood to bind its target DNA sequences only after IFNγ induction[43,44], and we previously detected no STAT1 binding sites in HeLa cells prior to induction[13]. In IFNγ-treated HeLa cells, we identified 1969 STAT1 binding sites (Table S4), with 85.8% of STAT1 binding sites occurring distal to Known Gene 5’ ends (Figure 3.3E). We observed that 447 (26.5%, $P = 1.47e^{-116}$) distal STAT1 binding sites overlapped enhancers that were predicted in HeLa cells prior to induction (Figure 3.3F), further supporting the accuracy and utility of our enhancer map and providing additional evidence that some enhancers may be poised for activation prior to binding by sequence-specific transcription factors (see Discussion).

3.8 Discussion

We have described the first genome-wide map of human transcriptional enhancers predicted on the basis of H3K4me1/H3K4me3 chromatin signatures. The distribution of predicted enhancers throughout the human genome is strongly related to cell type-specific gene expression patterns, and the presence of additional marks (H3K27ac, p300, TRAP220, and DHS) at many enhancers is also correlated with their role in gene expression patterns. The classes' different relationships to gene expression could explain the weak H3K4me3 observed at some of the "active" enhancers near HeLa-specific expressed genes, as chromosome looping brings enhancers and promoters into close proximity and results in apparent H3K4me3 enrichment at these enhancers.

Additionally, we discovered enrichment of many known transcription factor motifs in the predicted enhancers, and identified many novel sequence motifs that appear to be enhancer-specific. Further experiments are needed to establish the function of the novel motifs, but as several of the known motifs correspond to factors that have been demonstrated to bind the predicted enhancers in a variety of cell types, the motif data offer a very useful resource for additional experiments investigating patterns of activator-mediated gene expression in diverse cellular contexts. The enhancer map itself will also be of great utility in annotating the function of potential regulatory elements identified in other experiments, as demonstrated by the significant overlap of enhancer predictions with ER, p53, p63, and STAT1 binding sites in cells of various lineages. Additional experiments are needed to determine the full complement of transcription factors that are bound to these enhancers in HeLa cells and other cell types.

We provide substantial sequence-based and *in vivo* evidence for a subset of poised enhancers that do not appear to be strongly correlated with HeLa-specific gene expression, but correspond to transcription factor binding sites found in other cell lineages or physiological conditions, supporting an epigenetic memory mechanism for maintaining transcriptional regulatory potential at some enhancers during lineage specification and other cellular processes. Indeed, “poised” chromatin structure involving H3K4me1 has been reported at distal regulatory regions in the chicken lysozyme locus even prior to activation of lysozyme gene expression[45,46]. Perhaps H3K4me1 is a marker of active or poised chromatin conformation at human enhancers. In yeast, it has been shown that methylation of the H3K4 residue inhibits the association of silencer protein Sir3 with histone H3, thereby preventing heterochromatin formation at H3K4-methylated regions[47]. Further, Isw1, a protein involved in remodeling chromatin at promoters, preferentially recognizes di- and tri-methylated H3K4[48], suggesting a mechanism whereby enhancers are maintained in an active conformation by the euchromatic H3K4me1 marker, while H3K4me3 specifically denotes the 5’ ends of genes for recognition by protein complexes responsible for regulating expression at promoters.

It is striking that a presumably cell type-specific enhancer map is capable of predicting the potential function of such a large fraction of putative regulatory elements in other cell types and conditions. In our experiments, we discovered that the β-globin enhancer HS2 is one of several known enhancers marked by the H3K4me1 enhancer chromatin signature, presumably in a poised or dormant condition as the genes in this locus are not expressed in HeLa cells. Further, many of the novel STAT1-bound enhancers that we identified in this study and previously[13] are marked by enhancer chromatin signatures even prior to binding and activation by STAT1. These findings contribute additional evidence that certain subsets of enhancers actively participate in cell type-specific gene regulation at any given time, while other subsets are maintained in a poised or dormant state, awaiting activation under appropriate cellular circumstances.

3.9 Methods

Experimental procedures: HeLa, K562, GM06990, and IMR90 cells were obtained from ATCC and cultured under recommended conditions. Chromatin preparation, ChIP, DNA purification, and LM-PCR were performed as previously described, using commercially available antibodies (α-H3K27ac, Abcam ab4729; α-H3K4me1, Abcam ab8895; α-H3K4me3, Upstate 07-473; α-p300, Santa Cruz sc-585; α-TRAP220, Santa Cruz sc-5334; α-STAT1, Santa Cruz sc-345). ChIP samples were hybridized to the NimbleGen genome-wide tiling microarray set (NimbleGen Systems, Inc.) as previously described[28] and to custom condensed enhancer microarrays (NimbleGen Systems, Inc.) using standard methods. The condensed enhancer microarrays consisted of tiled 10 kb windows around each of 38716 primary predicted enhancers and standard controls. DNase-chip was performed and the data analyzed as previously described[9].

ChIP-chip data analysis and chromatin signature-based predictions: Data were analyzed using standard methods, and ChIP-chip targets for p300, TRAP220, and STAT1 were selected with the Mpeak program. Promoters and enhancers were predicted as previously described[13], with slight modifications to account for probe spacing on these array platforms. Enhancer predictions were considered "signature-verified" if their averaged H3K4me1 and H3K4me3 enrichment profiles on the condensed enhancer microarrays were sufficiently correlated to known enhancer chromatin signatures. K-means clustering, intersection analysis, and other computational comparisons (to UCSC Known Genes, CAGE tags, CpG islands, ChIP-chip target lists, etc.) of the prediction sets were performed as previously described[13].

Gene expression and entropy analysis: Gene expression in the various cell lines was analyzed using HGU133 Plus 2.0 microarrays (Affymetrix) as described[28]. Specificity of expression was determined using a function of Shannon entropy as described[35] and the top, middle, and bottom 1000 genes from this analysis were designated as HeLa-specific expressed, non-specific expressed, and HeLa-specific unexpressed genes, respectively, for evaluation of enhancer enrichment in the insulator-defined domains containing the promoters for these classes of genes (as in Figure 3.2A-B), where insulators were defined by CTCF binding sites[23]. When counting enhancers around these promoters (as in Figure 3.2C), we included all enhancers within 200 kb of a promoter as long as they were still within the same insulator-defined domain as described above. Random distributions were generated by averaging the enrichment profiles around promoters of 100 iterations of randomly selected enhancer sets equal in size to each class of enhancers in this analysis (EI/II, EIII, EIV).

Motif analysis: Enhancer regions were defined as 2 kb windows centered on each prediction, and promoter regions were defined as 1 kb windows upstream from annotated TSS. Promoters regions were excluded from enhancer regions; repeats, exons and transposons were excluded from both. Motif conservation in each region was evaluated relative to the genomes of opossum, tenrec, elephant, armadillo, cow, dog, rabbit, rat and mouse, extracted from UCSC Genome Browser and used with permission. The mammalian tree, along with branch lengths, was computed using DNAML (PHYLIP package)[49] with the F84 nucleotide model of evolution in ~500kb of randomly selected exon sequence. Known and novel motifs were discovered as previously described[38], with the primary difference that instances were not required to have perfect conservation and were considered conserved if they were found across a number of species spanning at least 50% of the total branch length of the mammalian tree [Stark A, Kheradpour P, and Kellis M, in preparation]. We ranked motifs based on their over-conservation, measured as the probability of observing an substantially increased number of conserved motif instances compared to that expected for motifs of identical composition, and selected all motifs with $P < 1e^{-3}$. We evaluated a motif's enrichment as its over-abundance, or the hypergeometric probability of observing a substantially increased number of occurrences in the intergenic and intronic regions of the human genome (regardless of evolutionary conservation) compared to motifs of identical composition, with a cutoff of P < 1e^{-3}.

Supplementary data for the microarray experiments has been formatted for the UCSC genome browser via http://bioinformatics-renlab.ucsd.edu/enhancer (user: enhancer, password: h3k4me1)

3.10 Acknowledgment

Chapter 3, in full, is a reprint of the material as it appears in a manuscript submitted to Nature in June 2007, “A Genome-wide Map of Human Transcriptional Enhancers.” Heintzman ND, Hon GC, Kheradpour P, Stark A, Ching KA, Stuart RK, Harp LF, Hawkins RD, Ching CW, Liu H, Zhang X, Green RD, Crawford GE, Kellis M, and Ren B. The dissertation author was the primary investigator and author of this publication.

NDH, BR, HL, XZ, and RDG designed the transcription factor and histone ChIP-chip experiments; GEC designed and performed the DNase-chip experiments; NDH, RKS, LFH, RDH, and CWC conducted the ChIP-chip experiments; NDH, GH, and KAC analyzed the microarray data; PK, AS, and MK performed the motif and conservation analyses; NDH, PK, MK, and BR wrote the manuscript.

3.11 Tables and Figures

Table 3.1: De novo motifs enriched in predicted enhancer regions.

Known Match score represents the shared information content between novel and known motif [38]. Over-conservation is calculated as the excess conservation of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. Enrichment is calculated as the over-abundance of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. Enhancer-specific motifs are those lacking significant promoter enrichment. All significance values are expressed as Z-scores, corresponding to the number of standard deviations away from the mean of a normal distribution.

Name	Consensus	Known Match (score)	Enhancer over-conservation Z-score (stdev)	Enhancer enrichment Z-score (stdev)	Promoter over-conservation Z-score (stdev)	Promoter enrichment Z-score (stdev)	Promoter depletion Z-score (stdev)
M01	VTGABTCRC	AP-1 (80%)	36.0	37.1	15.7		
M02	TAATTGM	NKX2-5 (88%)	36.0	4.2	14.8		
M03	MAAKGTCR	SF-1 (80%)	35.7	20.4	11.7		
M04	CTTTGAW	TCF-4 (97%)	32.2	8.4	12.9		
M05	YGANTYRGC		31.1	26.8	13.1		4.6
M06	GGAARTGA	STAT1 (88%)	29.5	13.1	21.4	8.3	
M07	TAATTAC	CHX10 (80%)	27.2	4.3	11.0		
M08	YTGGCNNNNNKYCMR	NF-1 (82%)	25.5	30.2	11.3		13.7
M09	YCATTAGY		25.4	5.0	9.2		3.8
M10	ATYWGTCR		23.8	14.0	6.5		
M11	RCATTCCA	TEF-1 (79%)	20.9	21.5	8.2		
M12	RACAGMTGK	TAL-1ALPHA/E47 (86%)	20.3		8.8		9.3
M13	CNTRGCAAC		18.9	5.6	21.7		
M14	AAACCACA	AML1 (86%)	18.6	13.1	5.8		3.7
M15	TGASGTCR	CREB (85%)	18.4	12.7	18.6	15.7	
M16	TAAWTTA	POU6F1 (78%)	15.7				3.3
M17	GCCARGAA		15.7	7.9	5.3		13.9
M18	CACNAGNGGG		15.5		8.3		
M19	GCTAWWWWTAG	MEF-2 (83%)	15.3	8.6	9.0	4.2	
M20	CATNANTAAT		15.1	5.2	5.8		
M21	TGTYKACR		14.6	3.3	6.8		
M22	GCCARNNNAAACA		12.0	15.1			
M23	TATTNNNNYYGGC		12.0	3.7			
M24	YGTCNRRACA		11.8	4.3			
M25	TAATGAGC	CHX10 (83%)	11.6		5.5		
M26	TAATTGGC	CHX10 (83%)	11.5		4.2		
M27	AGGTTAAT		11.5	3.7			
M28	ATTANNNNYGACR		10.5				
M29	GTCTAGAC		10.3	4.4	4.1		
M30	YGTCRNNNNNNATTA		10.3				
M31	CANYAGVTGGC		10.1		7.3		3.6
M32	YGTCRRTCA		9.8		9.7		
M33	SATCAATCR	PBX-1 (84%)	9.5				
M34	YGATTNRNTGC		9.5	4.1	7.8	4.7	
M35	AGGCNNNNNGCCAR		8.3	9.9	3.8		18.7
M36	GCCRRNNNNNNNATTA		7.5				8.6
M37	GGAAWTNCCC	P65 (94%)	7.4	5.5	4.5		
M38	CAKCTGGA	RP58 (85%)	7.3	4.9	4.4		12.9
M39	AGCAGCTGC	AP-4 (90%)	6.5		4.2		
M40	RCCATATGGY		4.7				8.4
M41	GTYNCCANRGNRAC		3.7		4.1		

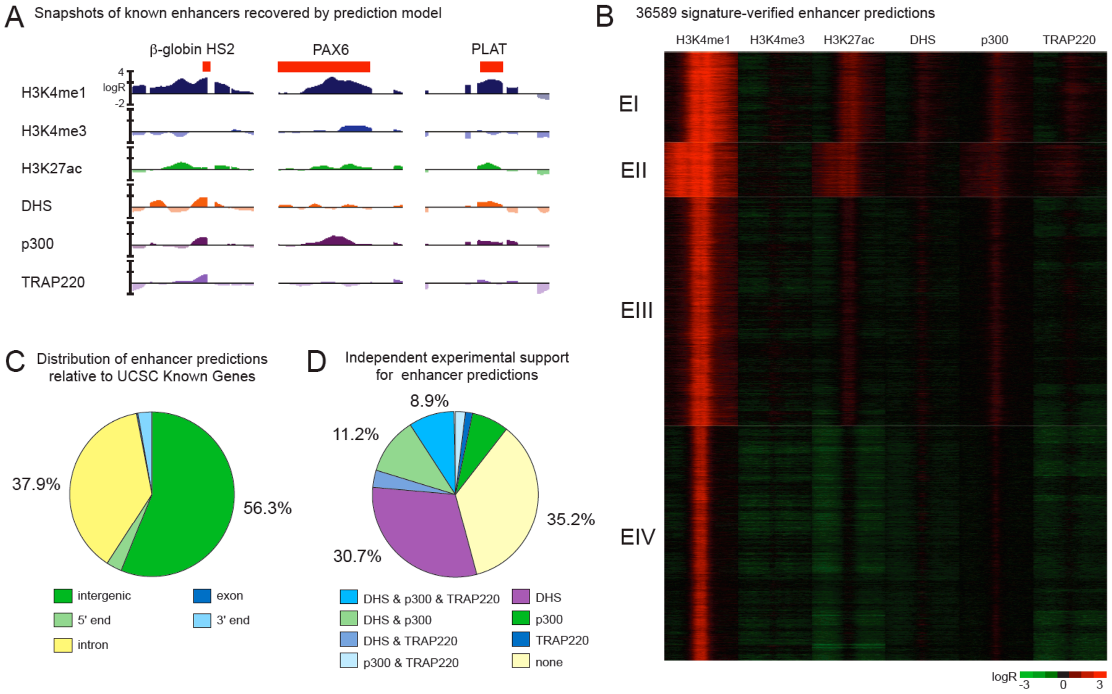

Figure 3.1: Enhancer predictions in the human genome.
(A) ChIP-chip enrichment profiles at several known enhancers (indicated in red) identified on the basis of their chromatin signature in HeLa cells: β-globin HS2 (chr11:5258371-5258665)[29], PAX6 (chr11:31630500-31635000)[30], PLAT (chr8:42191500-42192400)[31] (5 kb windows centered on enhancer predictions; images generated in part at the UCSC Genome Browser). (B) We predict 36589 enhancers in HeLa cells based on chromatin signatures for H3K4me1 and H3K4me3 as determined by ChIP-chip using genome-wide tiling microarrays and condensed enhancer microarrays (see text). Enhancers are clustered into four classes on the basis of histone modifications (H3K4me1, H3K4me3, and H3K27ac), DNaseI hypersensitivity (DHS), and binding of p300 and TRAP220. Classes are labeled EI-EIV to simplify discussion. (C) Most enhancers have intergenic (56.3%) or intronic (37.9%) localization relative to UCSC Known Gene 5' ends. (D) 64.8% of predicted enhancers are supported by DHS, binding of p300 or TRAP220, or combinations thereof.

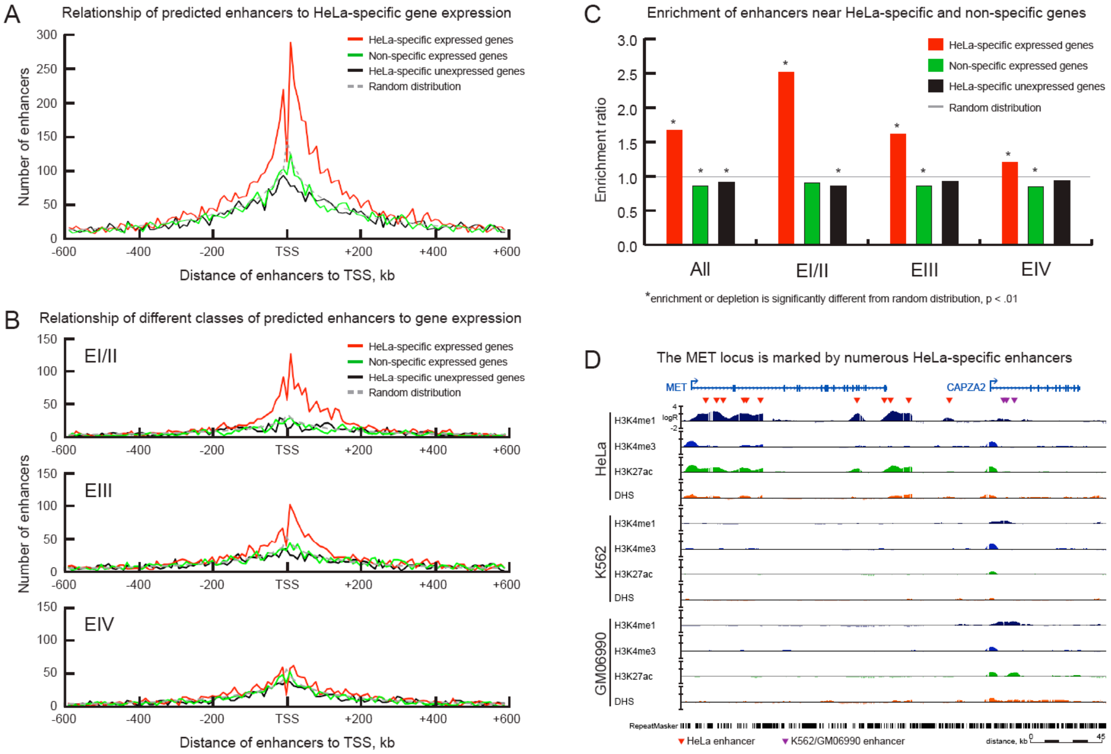

Figure 3.2: Enhancers are enriched around HeLa-specific expressed genes. (A) Enhancer localization relative to the promoters of genes that are HeLa-specific expressed compared to K562, GM06990, and IMR90 cells (red), non-specific expressed (green), HeLa-specific unexpressed (black), and a random distribution (dashed grey). Comparison of these different classes shows that enhancers are enriched around HeLa-specific expressed genes within insulator-defined domains. (B) Enhancer localization as in (A) but dividing the enhancers into classes EI/II, EIII, and EIV (as in Figure 1B), demonstrating that varying levels of H3K27ac, DHS, p300, and TRAP220 at enhancers are related to enhancer enrichment around HeLa-specific expressed genes. (C) We counted the number of enhancers within 200 kb of promoters (in the same insulator-defined domain) to quantify the enhancer enrichment at HeLa-specific expressed genes (red), non-specific expressed (green), HeLa-specific unexpressed (black), and random (grey), for all enhancers and for each class. (D) The MET gene is specifically expressed in HeLa cells and is markedly enriched in HeLa-specific enhancers (red triangles), while the CAPZA2 gene is expressed similarly in different cell types and lacks HeLa-specific enhancers (image generated in part at the UCSC Genome Browser).

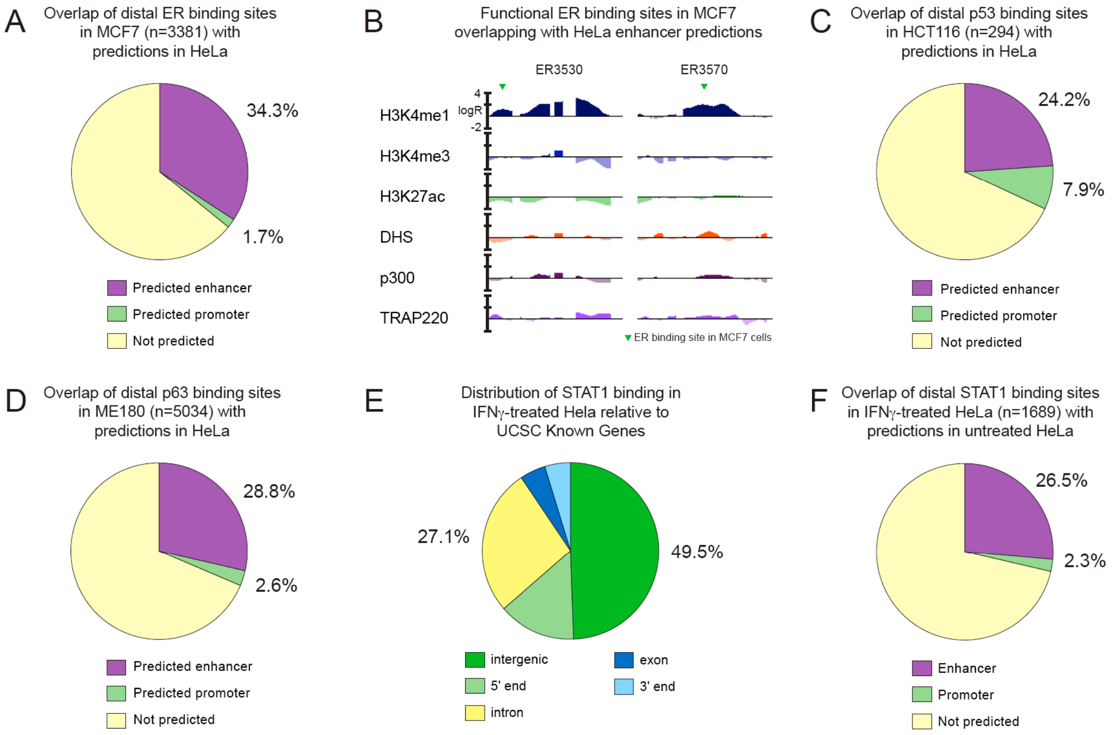

Figure 3.3: Enhancer predictions overlap with transcription factor binding sites in diverse cell lineages and conditions.

(A) 34.3% of promoter-distal ER binding sites in MCF7 cells overlap with enhancer predictions in HeLa cells. (B) ChIP-chip enrichment profiles at two ER binding sites with demonstrated function that display enhancer chromatin signatures (5 kb windows centered on enhancer predictions, ER binding sites[40] noted as green triangles; images generated in part at the UCSC Genome Browser). (C) 24.2% of promoter-distal p53 binding sites in HCT116 cells overlap with enhancer predictions in HeLa cells. (D) 28.8% of promoter-distal p63 binding sites in ME180 cells overlap with enhancer predictions in HeLa cells. (E) Genome-wide ChIP-chip identified 1969 STAT1 binding sites in HeLa cells treated with IFNγ, most of which are located distal to UCSC Known Gene 5' ends. (F) 26.5% of promoter-distal STAT1 binding sites in treated HeLa cells overlap with enhancer predictions in untreated HeLa cells.

Supplementary Figure and Table Legends
available at http://bioinformatics-renlab.ucsd.edu/enhancer

Figure S1: Active promoter predictions in the human genome.
(A) We predict 13116 active promoters in HeLa cells based on chromatin signatures for H3K4me1 and H3K4me3 as determined by ChIP-chip using genome-wide tiling microarrays. (B) 75% of promoter predictions map to 5' ends of UCSC Known Genes, indicating a high degree of specificity. (C) 76% of active promoters (defined as RefSeq TSS for expressed transcripts) are correctly predicted, indicating a high degree of sensitivity. (D) 85.1% of promoter predictions overlap with CpG islands (defined by UCSC Genome Browser), accounting for close to half of the CpG islands in the genome.

Table S1: Predictions of active promoters in the genome of HeLa cells.
Coordinates are listed in hg17 for 13116 active promoter predictions.

Table S2: Predictions of enhancers in the genome of HeLa cells.
Coordinates are listed in hg17 for 36589 signature-verified enhancer predictions.

Table S3: Known motifs in predicted enhancers.
Enrichment of motifs in enhancers was analyzed as previously described[38]. Over-conservation and Enrichment are calculated as the excess conservation and over-abundance, respectively, of a motif in enhancers or promoters relative to that expected for a random motif of identical composition. All significance values are expressed as Z-scores, corresponding to the number of standard deviations away from the mean of a normal distribution.

Table S4: STAT1 binding sites in the genome of HeLa cells.
Coordinates are listed in hg17 for 1969 STAT binding sites as determined by ChIP-chip.

3.12 References

1. Heintzman ND, Ren B. The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cell Mol Life Sci* 2007;64:386-400.

2. Maston GA, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genome. *Annu Rev Genomics Hum Genet* 2006;7:29-59.

3. Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* 2000;14:2551-69.

4. Nightingale KP, O'Neill L P, Turner BM. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr Opin Genet Dev* 2006;16:125-36.

5. Miller W, Makova KD, Nekrutenko A, Hardison RC. Comparative genomics. *Annu Rev Genomics Hum Genet* 2004;5:15-56.

6. Stone EA, Cooper GM, Sidow A. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu Rev Genomics Hum Genet* 2005;6:143-64.

7. de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* 2005;15:1061-72.

8. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 2006;444:499-502.

9. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* 2006;3:503-9.

10. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, Stamatoyannopoulos JA. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* 2006;3:511-8.

11. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* 1988;57:159-97.

12. Felsenfeld G. Chromatin unfolds. *Cell* 1996;86:13-9.

13. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007;39:311-8.

14. Kim TH, Ren B. Genome-Wide Analysis of Protein-DNA Interactions. *Annu Rev Genomics Hum Genet* 2006;7:81-102.

15. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. *Cell* 2007;129:823-37.

16. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;316:1497-502.

17. Barrera LO, Ren B. The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol* 2006;18:291-8.

18. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. *Cell* 2007;128:669-81.

19. Hawkins RD, Ren B. Genome-wide location analysis: insights on transcriptional regulation. *Hum Mol Genet* 2006;15 Spec No 1:R1-7.

20. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, Jurka J, Kamal M, Mauceli E, Searle SM, Sharpe T, Baker ML, Batzer MA, Benos PV, Belov K, Clamp M, Cook A, Cuff J, Das R, Davidow L, Deakin JE, Fazzari MJ, Glass JL, Grabherr M, Greally JM, Gu W, Hore TA, Huttley GA, Kleber M, Jirtle RL, Koina E, Lee JT, Mahony S, Marra MA, Miller RD, Nicholls RD, Oda M, Papenfuss AT, Parra ZE, Pollock DD, Ray DA, Schein JE, Speed TP, Thompson K, VandeBerg JL, Wade CM, Walker JA, Waters PD, Webber C, Weidman JR, Xie X, Zody MC, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer SC, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Bloom T, Boguslavskiy L, Bonnet C, Boukhgalter B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M, Dorje P. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* 2007;447:167-77.

21. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799-816.

22. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306:636-40.

23. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 2007;128:1231-45.

24. Kim TH, Barrera LO, Qu C, Van Calcar S, Trinklein ND, Cooper SJ, Luna RM, Glass CK, Rosenfeld MG, Myers RM, Ren B. Direct isolation and identification of promoters in the human genome. *Genome Res* 2005;15:830-9.

25. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes. *Bioinformatics* 2006;22:1036-46.

26. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;38:626-635.

27. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res* 2002;12:996-1006.

28. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. *Nature* 2005;436:876-80.

29. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* 2005;15:1051-60.

30. Kleinjan DA, Seawright A, Schedl A, Quinlan RA, Danes S, van Heyningen V. Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Hum Mol Genet* 2001;10:2049-59.

31. Wolf AT, Medcalf RL, Jern C. The t-PA -7351C>T enhancer polymorphism decreases Sp1 and Sp3 protein binding affinity and transcriptional responsiveness to retinoic acid. *Blood* 2005;105:1060-7.

32. Hatzis P, Talianidis I. Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol Cell* 2002;10:1467-77.

33. Wang Q, Carroll JS, Brown M. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol Cell* 2005;19:631-42.

34. Atchison ML. Enhancers: mechanisms of action and cell specificity. *Annu Rev Cell Biol* 1988;4:127-53.

35. Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ, Jr. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol* 2005;6:R33.

36. Rasola A, Fassetta M, De Bacco F, D'Alessandro L, Gramaglia D, Di Renzo MF, Comoglio PM. A positive feedback loop between hepatocyte growth factor receptor and beta-catenin sustains colorectal cancer cell invasive growth. *Oncogene* 2007;26:1078-87.

37. Tsai HW, Chow NH, Lin CP, Chan SH, Chou CY, Ho CL. The significance of prohibitin and c-Met/hepatocyte growth factor receptor in the progression of cervical adenocarcinoma. *Hum Pathol* 2006;37:198-204.

38. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 2005;434:338-45.

39. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, Wang Q, Bekiranov S, Sementchenko V, Fox EA, Silver PA, Gingeras TR, Liu XS, Brown M. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* 2006;38:1289-97.

40. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 2005;122:33-43.

41. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y. A global map of p53 transcription-factor binding sites in the human genome. *Cell* 2006;124:207-19.

42. Yang A, Zhu Z, Kapranov P, McKeon F, Church GM, Gingeras TR, Struhl K. Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol Cell* 2006;24:593-602.

43. Brivanlou AH, Darnell JE, Jr. Signal transduction and the control of gene expression. *Science* 2002;295:813-8.

44. Levy DE, Darnell JE, Jr. Stats: transcriptional control and biological impact. *Nat Rev Mol Cell Biol* 2002;3:651-62.

45. Kontaraki J, Chen HH, Riggs A, Bonifer C. Chromatin fine structure profiles for a developmentally regulated gene: reorganization of the lysozyme locus before trans-activator binding and gene expression. *Genes Dev* 2000;14:2106-22.

46. Lefevre P, Lacroix C, Tagoh H, Hoogenkamp M, Melnik S, Ingram R, Bonifer C. Differentiation-dependent alterations in histone methylation and chromatin architecture at the inducible chicken lysozyme gene. *J Biol Chem* 2005;280:27552-60.

47. Santos-Rosa H, Bannister AJ, Dehe PM, Geli V, Kouzarides T. Methylation of H3 lysine 4 at euchromatin promotes Sir3p association with heterochromatin. *J Biol Chem* 2004;279:47506-12.

48. Santos-Rosa H, Schneider R, Bernstein BE, Karabetsou N, Morillon A, Weise C, Schreiber SL, Mellor J, Kouzarides T. Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol Cell* 2003;12:1325-32.

49. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.* 2005.

Chapter 4

Functional characterization of genes by coupling RNAi library selection with DNA microarrays: RNAi Genomic Screen (RiGS)

Abstract

Much of our knowledge of gene function is derived from genetic studies using organisms such as *S. cerevisiae*, *C. elegans*, and *D. melanogaster*, in which mutations associated with specific phenotypes are easy to identify. A typical genetic approach involves mutagenesis of the genome, isolation of mutants, and identification of the affected genes. However, this standard approach is not amenable for genetic analysis of humans due to the difficulties associated with various steps of the traditional genetic analysis. In order to develop a general methodology for functional analysis of genes in human and other mammalian cells, we have coupled the recently developed RNA inference (RNAi) methods with DNA microarray technologies. Our method involves the construction of library of RNAi vectors that is stably introduced into the genomes of a population of cells, followed by quantification (using DNA microarrays) of each RNAi targeting vector's frequency in the population during a phenotypic selection. The microarrays we employ contain DNA sequences corresponding to all RNAi vectors and can reveal relative abundance of the RNAi vectors through hybridization of genomic DNA samples to the array. The representation frequency of each RNAi vector in the population allows one to infer the functional requirement of the target gene in a particular cellular process. In a small-scale study, we have proven

the sensitivity and specificity of this method, which can be easily scaled-up for efficient functional analysis of the entire human genome.

4.1 Introduction

RNA interference (RNAi) has emerged as a versatile tool for silencing a specific gene and analyzing its function in a biological context[1-3]. Discovered as a natural defense mechanism in plants and extended to invertebrate and mammalian systems, RNAi utilizes cellular machinery to target and degrade mRNA transcripts in a highly specific fashion. Briefly, short (19-21 nt) RNA oligonucleotides bind to homologous mRNA sequences within the cell and direct them to the RNA-induced silencing complex (RISC), where the mRNA is degraded. The RNA oligonucleotides can come from long double-stranded RNAs (dsRNA) or small interfering RNAs (siRNA) injected into the cell, or short hairpin RNAs (shRNA) that are transcribed endogenously under the control of an RNA-polymerase III-driven promoter. In any case, RNAi results when DICER processes the dsRNA, siRNA, or shRNA into a 19-21 nt oligo capable of targeting a specific mRNA transcript for RISC-mediated degradation. This phenomenon has been an area of active study since its discovery, and our understanding of its mechanisms continues to improve.

We propose a method to combine RNAi and microarray technology into a high-throughput functional genetic screen. Our technique, RNAi Genomic Screen (RiGS), allows genome-wide analysis and identification of genes functionally involved in a host of biological systems, implicating known genes and discovering novel players in a single experiment. This method involves the construction of a library of RNAi vectors, introduction of the RNAi library to host cells and phenotypic selection of cells, extraction of RNAi vectors from the cells by polymerase chain reaction (PCR), and quantification of RNAi vector frequency by hybridization to DNA microarrays.

4.2 Description of the RiGS method

4.2.1 Construction of RNAi vector library

Our method utilizes a retroviral system to stably express shRNAs against a panel of genes in target cells (Figure 4.1). Oligos (80 nt) are designed to contain RNA polymerase III transcriptional start and stop sites flanking a 19 nt RNAi target sequence and its reverse complement separated by a short (9nt) spacer. Oligos encoding shRNAs against multiple genes are pooled together and the complementary strand of each oligo is synthesized in batch. First, a common 20nt primer is annealed to the 3' end of the 80-mer oligos, and Bst DNA polymerase enzyme is used to synthesize and extend the duplex. The resulting double-stranded 80-mer DNAs are gel purified and digested with restriction enzymes to generate BamHI and HindIII overhangs. The digested small inverted repeat DNA is ligated into a retroviral vector, pSUPER-retro, downstream of an H1 RNA polymerase III promoter with the recipient BglII and HindIII overhangs. Chemically competent E. coli cells are transformed using the ligated DNA and plated onto a large 500cm^2 LB agar plates containing 100ug/mL ampicillin. The entire collection of colonies on the plate is harvested and their plasmid DNA extracted. The purified plasmid DNA contains a library of uniformly represented RNAi constructs. Transfection of this library into a packaging cell line (bosc23) results in a retroviral RNAi library capable of infecting a target cell population and knocking down a single gene in each infected cell.

4.2.2 Stable introduction of RNAi library to host cells

A target cell line is infected with the retroviral RNAi library and allowed to grow for time sufficient for stable integration of the RNAi vector, expression of the shRNA, and resultant knock down of target gene expression. The infected population is then subjected to some selective condition (differentiation, proliferation, drug response, apoptosis, etc.) and separated on the basis of a phenotypic characteristic reflective of the selective condition (i.e., the appearance of a cell surface marker for differentiation, annexin-positive staining for apoptosis, etc.).

4.2.3 Isolation of RNAi vectors from genomic DNA of the cells

Genomic DNA is isolated from each subpopulation and from an appropriate control population. Due to the design of the retroviral RNAi construct, each cell's genome contains an integrated shRNA expression cassette with a sequence unique to the gene it knocks down, as well sequences common to all constructs; thus, the shRNA-encoding sequence not only results in reduced gene expression, but also serves as a unique tag representative of the target gene. Primers are designed to anneal to common sequences flanking the shRNA-encoding region such that PCR of genomic

DNA from the infected cells amplifies the shRNA-encoding region and yields a pool of short DNA fragments, each containing the unique shRNA tag specifying which gene has been targeted and knocked down.

4.2.4 Quantification of RNAi vector prevalence in the population using microarray

These shRNA tags are present in proportions representative of the number of knock down cells in each subpopulation, and when the PCR product is labeled and hybridized to a microarray, we can visualize enrichment or depletion of a particular shRNA tag in selected cells relative to control cells, indicating functional involvement of that target gene in the system under investigation. In such a manner, novel genetic players can be implicated in a variety of biological processes.

For example, consider a population of progenitor cells that can be induced to differentiate by an external stimulus. If these cells are infected with our retroviral RNAi library, then induced to differentiate and assayed by RiGS, genes necessary for differentiation would be identified by enrichment of their shRNA tags in the undifferentiated cells and concurrent depletion of these tags in the differentiated cells, as knocking down these genes prevents differentiation and retains cells in the undifferentiated state. Conversely, genes that inhibit differentiation would display opposite patterns of enrichment and depletion.

4.3 Results

We performed a “Color Test” to examine the sensitivity and specificity of our method in human embryonic kidney 293 (HEK293) cells stably expressing GFP (HEK293GFP). We infected these cells with a small retroviral library of 12 RNAi vectors targeting a total of 6 genes, including myc, max, p53, K-ras^{V12}, dsRed, and GFP. We separated the infected HEK293GFP cells into GFP-positive and GFP-negative populations using FACS, then extracted the genomic DNA from these two populations and from an unsorted control population. Using PCR with fluorescently-labeled common probes, we amplified and labeled the shRNA tags inserted into the host genome (DNA from sorted cells was labeled with cy5, unsorted control cell DNA with cy3) and hybridized the PCR products to a DNA microarray containing short oligonucleotide probes corresponding to the 12 RNAi target sequences present in the shRNA tags.

The results of triplicate experiments are shown in Figure 4.2. The RiGS Color Test shows a 5.5-fold enrichment of the shRNA targeting GFP (si/GFP) in cells with reduced fluorescence (GFP-) relative to the unsorted control population, while cells maintaining normal GFP expression (GFP+) showed a concurrent 2-fold depletion of si/GFP. No other shRNA tag showed more than 1.5-fold enrichment or depletion in either sorted subpopulation. Our results clearly display the sensitivity and specificity of RiGS in identifying the functional requirement of the GFP gene in the fluorescence of our cell line.

4.4 Discussion

RiGS is a powerful tool to identify novel genetic factors with functional roles in a variety of biological processes. Our initial goal with this assay is to design RNAi libraries and microarrays for analyzing the functional roles of all transcription factors in the human and mouse genomes. We will subsequently increase the scope of our assay to examine elements such as signaling pathways, and ultimately expand to include an entire genome in our libraries and microarrays. This will not only identify novel factors and novel roles for known factors in a given process, but will also be useful in identifying a complete cast of characters involved in combinatorial control expression of any gene. By using the native promoter of the gene to drive expression of a reporter, for example, we can identify factors required for regulating expression of the gene of interest.

Further refinements could include the use of an inducible system for an added level of control, as expression of interfering RNA species can be controlled by the presence or absence of a chemical agent acting on a responsive element in the promoter of the shRNA-expression construct. The RiGS method provides a means to identify and analyze genetic factors in a variety of biological systems, including human disease and drug response, and is a useful new tool for answering age-old biological questions, as has been demonstrated by other similar technologies[4-7].

4.5 Acknowledgment

RiGS was conceived and designed in 2003 by Bing Ren and Nathaniel Heintzman. Some modifications to the methodology, as outlined in section *4.2.1 Construction of RNAi vector library*, were designed by Tae Hoon Kim.

4.6 Figures

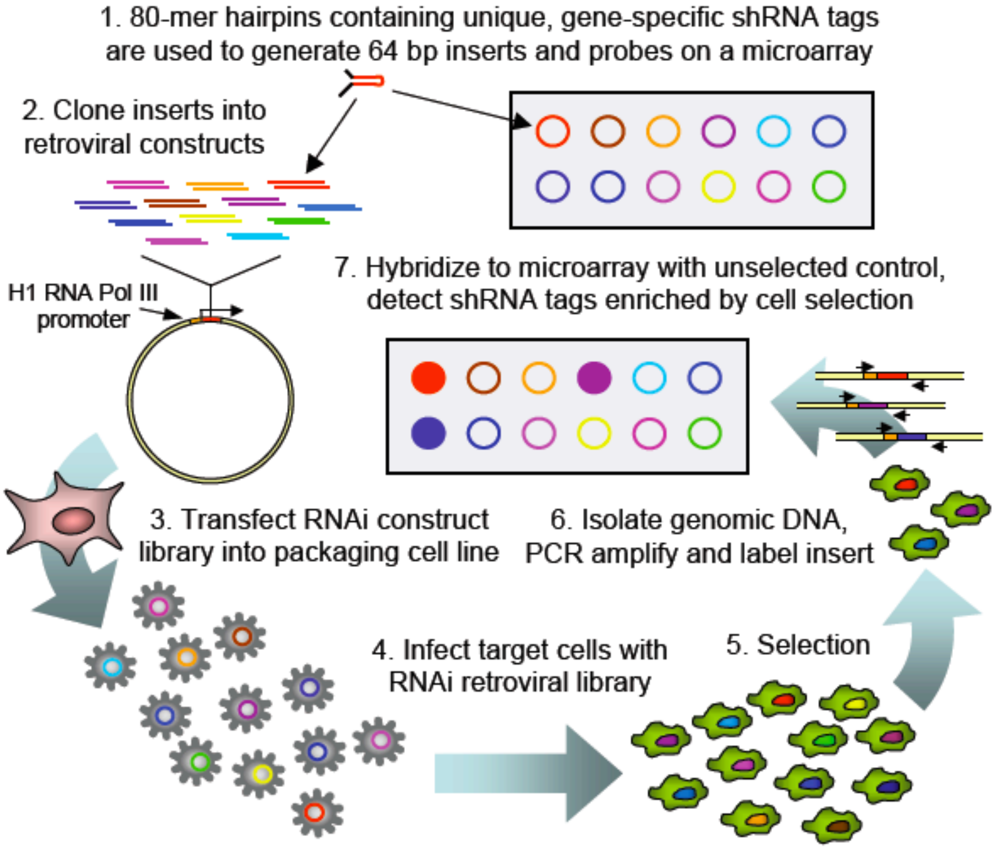
1. 80-mer hairpins containing unique, gene-specific shRNA tags are used to generate 64 bp inserts and probes on a microarray
2. Clone inserts into retroviral constructs
H1 RNA Pol III promoter
7. Hybridize to microarray with unselected control, detect shRNA tags enriched by cell selection
3. Transfect RNAi construct library into packaging cell line
6. Isolate genomic DNA, PCR amplify and label insert
4. Infect target cells with RNAi retroviral library
5. Selection

Figure 4.1: Schematic representation of RNAi Genomic Screen (RiGS)
For additional details of each step of this method, please refer to the corresponding section(s) of the text.

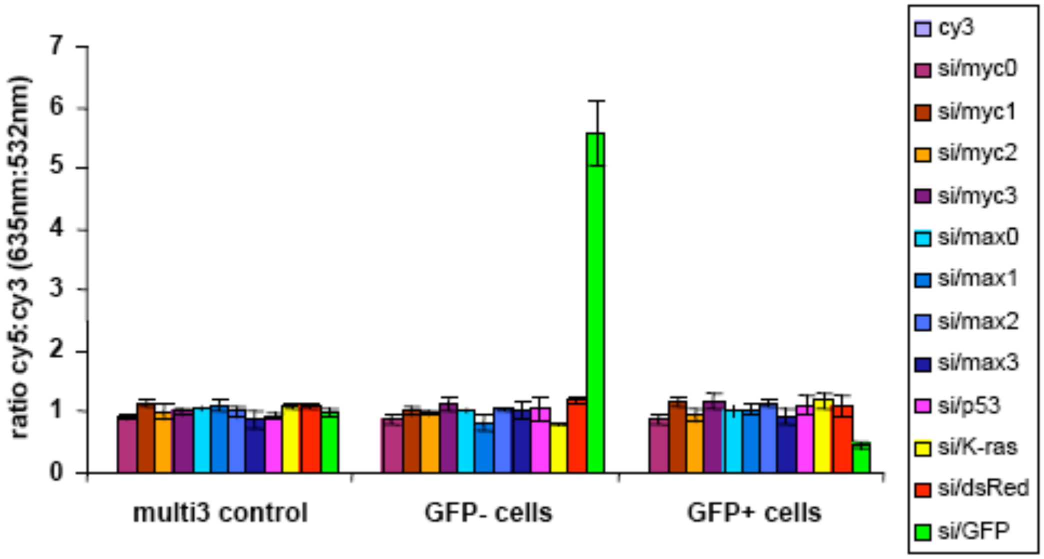

Figure 4.2: RiGS Color Test in HEK293GFP cells
HEK293 cells stably expressing GFP were infected with a small retroviral RNAi library including shRNAs against myc, max, p53, K-ras, dsRed, and GFP. Ten days after infection, cells were separated by FACS based on GFP expression. Genomic DNA was isolated from sorted and unsorted cells and the shRNA inserts were amplified and labeled by PCR and hybridized to a microarray; sorted DNA samples were labeled with cy5, unsorted with cy3. Examination of the cy5:cy3 ratio for each shRNA corresponding probe on the microarray shows a 5.5-fold enrichment of si/GFP in cells with reduced fluorescence (GFP-) relative to the unsorted control, while cells maintaining normal GFP expression (GFP+) showed a concurrent 2-fold depletion of si/GFP. No other shRNA tag showed more than 1.5-fold enrichment or depletion in either sorted subpopulation. Results are from triplicate experiments.

4.6 References

1. Chang K, Elledge SJ, Hannon GJ. Lessons from Nature: microRNA-based shRNA libraries. *Nat Methods* 2006;3:707-14.

2. Hannon GJ. RNA interference. *Nature* 2002;418:244-51.

3. Hannon GJ, Rossi JJ. Unlocking the potential of the human genome with RNA interference. *Nature* 2004;431:371-8.

4. Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K, Chang K, Westbrook T, Cleary M, Sachidanandam R, McCombie WR, Elledge SJ, Hannon GJ. A resource for large-scale RNA-interference-based screens in mammals. *Nature* 2004;428:427-31.

5. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, Agami R, Ge W, Cavet G, Linsley PS, Beijersbergen RL, Bernards R. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* 2004;428:431-7.

6. Brummelkamp TR, Berns K, Hijmans EM, Mullenders J, Fabius A, Heimerikx M, Velds A, Kerkhoven RM, Madiredjo M, Bernards R, Beijersbergen RL. Functional identification of cancer-relevant genes through large-scale RNA interference screens in mammalian cells. *Cold Spring Harb Symp Quant Biol* 2004;69:439-45.

7. Brummelkamp TR, Fabius AW, Mullenders J, Madiredjo M, Velds A, Kerkhoven RM, Bernards R, Beijersbergen RL. An shRNA barcode screen provides insight into cancer cell vulnerability to MDM2 inhibitors. *Nat Chem Biol* 2006;2:202-6.