UNIVERSITY OF CALIFORNIA, SAN DIEGO

Genome-wide mapping and analysis of mammalian promoters

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

Doctor of Philosophy

in

Bioinformatics

by

Leah Ortiz-Luis Barrera

Committee in charge:

Professor Bing Ren, Chair
Professor Philip E. Bourne, Co-Chair
Professor Vineet Bafna
Professor James T. Kadonaga
Professor Wei Wang

2007

The dissertation of Leah Ortiz-Luis Barrera is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2007

To my Mom, Ate, Mia, and Joseph
In memory of Dad and Lola

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY DATA TABLES

ACKNOWLEDGMENTS

This dissertation would not possible without the guidance of Prof. Bing Ren. As an advisor he has been energetic, approachable, and a source of keen insights. Bing has also facilitated many opportunities for collaboration and training – from my involvement in the ENCODE project to my selection as a workshop instructor for the CSHL Systems Biology Workshop. I am also grateful to all my previous research mentors who have contributed to my scientific training – from high school through my undergraduate years.

I would also like to thank my committee members: Professors Vineet Bafna, Phil Bourne, Jim Kadonaga, and Wei Wang. Their suggestions and discussion at committee meetings have been instrumental in improving my dissertation. The Bioinformatics Program faculty, in particular Shankar Subramaniam and Steve Wasserman, gave critical advice in my first year – suggesting laboratories to join and encouraging me to apply for fellowships. I would like to thank Prof. Glenn Tesler for demonstrating a deep concern for teaching quality. It was a pleasure learning from him as a TA. Finally, I would like to thank Dr. Yingyao Zhou for supervising a productive internship at GNF.

Dr. Ming Zheng and Prof. Yingnian Wu were remarkable collaborators with inspiring expertise in statistics and computation. Dr. Andrew Smith taught me the finer points of motif-finding and gave me an appreciation for rigor in this field by his example.

I thank colleagues in the Ren Lab who have made my graduate training more pleasant and productive. In particular I would like to thank Dr. Tae Hoon Kim and Dr. Zirong Li for the fruitful collaborations in which their experimental expertise and biological insights were indispensable. Dr. Chunxu Qu and Dr. Keith Ching have been lifesavers as bioinformatics staff scientists in the lab. Sara Van Calcar and Rhona Stuart

were patient and lively teachers of ChIP-chip and other experimental procedures. Nate M and Saurabh have been good "neighbors" for threshing ideas and discussion. Finally, the lab as a whole – Nate H, Gary, Lindsey, Zhen, Eugene, Christina, Leonard, and Esther – for their collegiality and the Krishna food get-togethers on Wednesdays.

I thank colleagues in the Bioinformatics program, in particular, Silpa Suthram and Thuy Vo, for their friendship and research insights. I am lucky to enter in the same class with talented students whose success has inspired me throughout graduate school. I would also like to highlight, Chris Benner and Kristine Briedis, for providing avenues for recreation by organizing IM sports every year. Finally I would like to thank "younger" students like Mary Pacold who forced me to reckon with a role as mentor by seeking my research advice.

I would like to thank friends and family who have kept me happy and well-fed throughout this challenging journey: Barbara Mattson, Sandi Sherman, and Ping Wang for the many dinners, coffees, cakes, and pies; Grace Liu for music and madness; All my aunts and uncles, Barrera and Ortiz-Luis, for their ready support at all points of my life; My Dad, for the inspiration to do well that endures beyond his passing; My mom and my sisters, whose love, support, and tireless care has made me feel blessed throughout my life; My fiancé Joseph, for inspiration.

I am grateful to the Ford Foundation for a 3-year pre-doctoral fellowship and to Dr. Patrick Mang for pointing out this funding source.

Chapter 1, in part quotes sections from: Barrera, Leah O.; Ren, Bing. The transcriptional regulatory code of eukaryotic cells, Current Opinion in Cell Biology, Vol. 18, 2006. I was a primary author of this review.

Chapter 2 in part quotes sections from the following publications: (1) Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press. (2) Zheng, Ming; Barrera, Leah; Ren, Bing; Wu, Yingnian. ChIP-chip: data, model, and analysis. Biometrics, in press. I was a a secondary author and researcher in these works. For (1), I wrote the sections dealing with ChIP-chip data analysis. For (2), I contributed to the development of the model, testing of the algorithm, and in editing the manuscript.

Chapter 3, is a reprint in full of the material as it appears in: Kim, Tae H; Barrera, Leah O; Zheng, Ming; Qu, Chunxu; Singer, Michael A.; Richmond, Todd A.; Wu, Yingnian; Green, Roland; Ren, Bing. A high-resolution map of active promoters in the human genome. Nature, Vol.436, 2005. I was a primary co-author and researcher of this work.I performed the bulk of the computational portion and analysis of the research. The other primary co-author performed all the experimental assays and validation. Other co-authors of the paper supervised and directed the work or contributed analytical tools (Mpeak) and experimental materials (NimbleGen Arrays).

Chapter 4, in full is a manuscript prepared for submission as: Barrera, Leah O.; Li, Zirong; Smith, Andrew D; Zhang, Michael Q; Green, Roland; Ren, Bing. Genome-wide promoter profiling of mammalian tissues. I was a primary co-author and researcher of this work. I performed the computational portion and analysis of the research and wrote the paper. The other primary co-author performed all the experimental assays and validation. Other co-authors of the paper supervised and directed the work or contributed analytical tools (DME *de novo* motif finder) and experimental materials (NimbleGenArrays).

VITA

2002 B.S, Stanford University (Stanford, CA)

2007 Ph.D., University of California, San Diego (La Jolla, CA)

PUBLICATIONS

1. **Barrera, LO***, Li Z*, Smith A, Zhang MQ, Green R, Ren,B. "Genome-wide mapping and analysis of tissue-specific promoters. Submitted
2. Kim TH, **Barrera LO**, Ren B. Genome-wide analysis of protein-binding in mammalian cells. *Current Protocols in Molecular Biology*, in press.
3. Zheng M, **Barrera LO**, Ren B, Wu Y. ChIP-chip: Data, Model, and Analysis. *Biometrics*, in press.
4. ENCODE Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, accepted.
5. Heintzman 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. *Nature Genetics* 2007, 39:311-318.
6. **Barrera LO**, Ren B. The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. *Current Opinion in Cell Biology* 2006, 18:291-298.
7. 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, 436:876-880.
8. 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 Research* 2005, 15:830-839.
9. **Barrera L**, Benner C, Tao YC, Winzeler E, Zhou Y: Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays. *BMC Bioinformatics* 2004, 5:42.

(*) Authors contributed equally to the work

ABSTRACT OF THE DISSERTATION

Genome-wide mapping and analysis of mammalian promoters

by

Leah Ortiz-Luis Barrera

Doctor of Philosophy in Bioinformatics
University of California, San Diego, 2007

Professor Bing Ren, Chair
Professor Philip Bourne, Co-Chair

Mammalian organisms such as mouse and human are characterized by large genomes of 2-3 billion base pairs. Sequencing of these genomes has revealed that only a small fraction, ~1.5%, encodes protein-coding genes. The diversity of more than 200 cell types which make up mammals, from the zygote to the differentiated cell types which perform the functions of organs, is brought about by the coordinated expression of specific subsets of these genes. Control of gene expression is, in turn, mediated by the binding of transcription factors at non-coding genomic regulatory sequences such as promoters, enhancers, and insulators. Unraveling the control of gene expression, resulting in mammalian cell type diversity, thus entails the accurate and systematic characterization of these sequences.

In this work, we describe a pilot application of chromatin immunoprecipitation with microarrays (ChIP-chip) to define active promoters in human fibroblast cells. To do this, we mapped the genomic location of components of the transcription pre-initiation complex (PIC) using microarrays tiling the entire non-repetitive human genome sequence at 100 bp resolution. The scale and novelty of this high-throughput strategy entailed significant bioinformatics challenges. In particular, we highlight our model-based approach for the accurate identification of binding sites from the data. Interestingly, this pilot identification of 10,567 active promoters revealed the extent of alternative promoter usage within a single cell type, clustering of active promoters, and classes of genes based on PIC binding and transcript expression level.

We then extended our genome-wide promoter mapping strategy to characterize active promoters in mouse embryonic stem cells (mES) and adult organs. We mapped ~24,000 promoters across these samples, including 5,153 sites validating cap-analysis of gene expression (CAGE) 5' end data in addition to 16,976 annotated mRNA 5' ends. To profile promoter usage across tissues by relative occupancy of RNA polymerase II (Pol II), we adapted a quantitative index of tissue-specificity and thus overcome limitations of "bound" or "unbound" classification. We examined the sequence and epigenetic features of tissue-specific promoters defined by this measure and discovered a subset of promoters with enriched Pol II binding in mES persistently marked by H3K4me3 in adult tissues.

Chapter 1

Introduction

The complement of hereditary information, which differentiates a human from a mouse or a mouse from a fly, is encoded in a genome packed into every cell that makes up these organisms. The range of proteins which perform the biological processes and form the structures distinguishing the variety of cell types within these organisms are encoded as "genes" in the genome. The concept of the genome as information is due to its simplified representation as a long sequence specified by a four-letter alphabet (A, T, G, C). Three-letter combinations or codons map to 20 distinct fundamental units of proteins called amino acids and constitute a genetic code. Mammalian organisms such as mouse and human are characterized by large genomes from ~2-3 billion "letters" in total length. Sequencing of these genomes has revealed that only a small fraction (~1.5%) of the total length encodes protein-coding genes in these organisms [1,2].

Despite the simplified four-letter representation, a genome is not merely sequence. Within cells, the genome exists as double helical chains of deoxyribonucleic acid (DNA). The letters (A, T, G, C) correspond to nitrogen-containing bases (adenine, thymine, guanine, cytosine) that differentiate the fundamental units called nucleotides which make up each DNA chain. In living cells, the genome is dynamic because of the changing biochemical interactions between DNA and proteins. For instance, in eukaryotes, the genome exists as chromatin – a complex of proteins called histones along with DNA[3]. Changes in the compaction of and structure of chromatin, histone variant

composition and chemical modifications of histones can affect the accessibility of the associated DNA to other proteins which decode the underlying information. Broadly termed transcription factors, these proteins bind short DNA sequences and in turn affect the rate that specific genes are transcribed from the genome.

Although significant advances in characterizing the protein-coding gene content of mammalian genomes has been achieved through knowledge of the genetic code and sequence analysis, we are far from a complete characterization of the larger fraction of non-coding genomic sequence. Methods for mapping protein-DNA interactions *in vivo* have emerged as a powerful complementary strategy for characterizing the function of DNA elements other than protein-coding genes in the genome. These functional elements can be defined by their characteristic protein markers such as transcription factors, histone variants, or by associated histones with distinctive modifications.

In this dissertation, we demonstrate the feasibility of a genome-wide approach for characterizing a specific class of functional elements by mapping protein-DNA interactions. We describe the experimental method and technological developments that enabled this approach. The novelty of the method and the large volume of resulting genome-scale data entailed significant bioinformatics challenges. In particular, we highlight the development of a model to accurately predict protein-DNA interaction sites from the data. Largely, we reveal how these pilot applications of the genome-wide approach to map promoters in one mammalian cell type and in a comparative study across a panel of mammalian tissue types contribute to our understanding of a dynamic genome.

1.1 Control of Gene Expression

Despite the availability of mammalian genome sequences and the presumed exhaustive annotation of protein coding genes, our understanding of the mechanisms by which subsets of genes are expressed in a cell type or tissue-restricted manner remains quite limited [1,4-6]. Although assembly of eukaryotic DNA into chromatin confers additional levels of regulatory control, the general model of transcriptional activation suggests the critical role of sequence-specific DNA-binding factors at regulatory sequences such as enhancers and promoters in activating transcription initiation by RNA polymerase II and the general transcription machinery at core promoters [7,8]. Thus, the development of systematic genome-scale approaches toward the characterization of these enhancers and promoters critical for mediating tissue-specific expression of linked transcriptional loci will represent key advances in our attempts to understand this process.

1.1.1 Regulatory Elements in Gene Activation

Previous investigations of tissue-specific expression have been mainly carried out on a gene-by-gene basis and directed toward the elucidation of three types of regulatory elements known to mediate tissue-specific transcription efficiency (1) chromatin openers, (2) enhancers, (3) promoters[9]. Chromatin openers affect the decondensation of repressed chromatin to a potentially active state, and thus increase the accessibility of the gene(s) within the locus to transcription machinery. A related class of *cis*-acting element known as the locus control region (LCR) has been characterized for a number of tissue-specific genes and gene loci using DNAse-I hypersensitivity site mapping. Canonical examples include the β-globin LCR which mediates erythroid cell-specific expression of globin

genes and the T-cell-specific TCR-α/δ LCR. Aside from containing chromatin opening elements, these LCRs are also characterized as containing cell-lineage specific enhancer activity[10]. Classical enhancers by definition stimulate transcription efficiency independently of orientation and distance by the binding of their cognate transcription factors. These chromatin openers and enhancers are commonly defined by tissue-specific DNAse-I hypersensitivity sites[11]. Their tissue-restricted activity and ability to act at a distance makes them particularly challenging to identify and match to their affected transcriptional loci. A promoter on the other hand, has been less challenging to define once tissue-specific expression of a transcript or gene is established.

1.1.2 Promoters

Promoters are typically characterized as containing two distinct structures: (1) a core promoter region that binds the general transcription initiation machinery defined as approximately [-35, +35] bp over the transcriptional start site, and (2) a proximal promoter region containing binding sites for sequence-specific DNA-binding proteins, similar to those contained within enhancers, which can activate or repress transcription initiation efficiency [4,11,12] (Figure 1-1).

Core promoters are thought to be generally inactive *in vivo* without transcriptional activation mediated through co-activators. These co-activators bridge the effects of sequence-specific transcription factors bound to short sequence motifs at the proximal promoter or an enhancer [11,13,14]. Although recent studies suggest that cell-type specific components of the general transcriptional machinery can contribute to promoter selectivity leading to tissue-specific gene expression, core promoters are considered to

have limited tissue or cell-type specificity[15]. Nonetheless, detailed biochemical studies have revealed the importance of short sequence motifs characterizing core promoters such as the TFIIB recognition element (BRE), TATA box, Initiator (INR), Motif Ten Element (MTE), and the Downstream Core Promoter Element (DPE) [12,16-18]. The variable combinations of these short sequence motifs at core promoters are suggested to contribute to the complex control of gene regulation[19,20].

Proximal promoters on the other hand have been characterized to contain short sequence motifs that are sufficient to direct tissue-specific transcription in transient transfection studies[21]. One of the earliest well characterized examples is the proximal promoter of the albumin gene. This highly-expressed gene in liver was shown to contain cognate binding sites for liver-enriched factors such as HNF-1 and C/EBP within 150 bp upstream of the transcriptional start site (TSS)[9]. The role of promoters in the complex control of gene expression is substantiated by the degree to which complex promoter structures are conserved among species[22]. Sequence comparison of human and mouse genes have shown homologous block structures in promoters with regions of conservation extending on average up to 510 bp upstream of the TSS[23]. More recently, discovery of known and novel conserved sequence motifs linked to tissue-specific expression in human from a comparison of promoters in several mammals clearly supports the idea that motifs in proximal promoters can be predictive of tissue-specific expression[24].

1.1.3 Tissue-Specific Expression

Clearly directed genome-scale approaches toward characterization of tissue-specific

promoters will be critical in the task of annotating tissue-specific expression of transcripts and in beginning to understand mechanisms of the process. Most genome-scale approaches to date have been accomplished by combining information for tissue-specific expression of genes from expression profiling experiments with computational strategies for identifying possible motifs for transcription factor binding sites. One computational study has described a method for systematic identification of tissue-specific transcription factor binding sites by focusing on sequence motif differences in the promoters of differentially expressed genes across tissue types. This study used manually curated sets of tissue-specific genes to define motifs for transcripts with liver-enriched and muscle-enriched expression[25]. More general sequence analyses of promoter features related to tissue-specificity from microarray-based and expressed sequence tag expression data thus far have mainly identified the general correlation of tissue-specific genes with promoters that contain TATA boxes and lack CpG islands, while the least tissue-specific genes are generally correlated to CpG islands[26,27]. Aside from general sequence features, expression profiling studies of known and predicted genes across a panel of 79 human and 61 mouse human tissues have identified hundred of regions of correlated transcription (RCT), linearly co-localized genes with similar expression patterns across tissues, including some that are subject to tissue-specific expression. These tissue-specific RCTs were posited to be regulated by common promoter elements or through higher-order gene regulation leading to site-specific remodeling of chromatin to active domains, but both hypotheses have yet to be substantiated by further analyses[28]. A key limitation of these studies dependent on expression data is that the analyses focus on genes rather than on transcripts, which different promoters from the same gene could

generate. The rate of occurrence of alternative transcript start sites has been estimated to range from 9% to 52% of known genes in mouse [21,27,29,30]. Increasing evidence for the prevalence of alternative promoter usage and their role in mediating tissue-specific expression in mammals, thus further underscores the importance of the accurate identification of core and proximal promoter sequences linked to tissue-specific transcript expression [22,31].

1.2 Mapping Protein-DNA Interactions in the Genome

The developments of tools for large-scale mapping of in vivo protein-DNA interactions, such as chromatin immunoprecipitation with microarrays (ChIP-chip) are enabling global views of transcription factor binding and chromatin context. Applications of these strategies to characterizing genomic DNA interactions with histone proteins, general transcription factors and sequence-specific transcription factors are beginning to identify regulatory regions, unravel the chromatin features at specific types of regulatory regions, and reveal the coordinated roles of transcription factors and chromatin modifications in the control of gene expression in eukaryotic genomes. Prior to the publication of the work described in this dissertation, high-resolution genome-scale maps of protein-DNA interactions were completed only in yeast. Protein-DNA maps in flies or mammals surveyed selected regions or chromosomes. In the following sections, we give a brief background on chromatin, an overview of ChIP-chip, and review the advances based on this strategy prior to our work.

1.2.1 Chromatin

In eukaryotes, genomic DNA is partitioned among chromosomes within the nucleus in each cell, and each chromosome consists of a long stretch of DNA and proteins, referred to as chromatin. There are various levels of chromatin condensation leading to the densely packed chromosome[3]. The nucleosome, the fundamental unit of chromatin organization, consists of 146bp of genomic DNA spooled in less than two turns around a disk-like histone octamer. Each histone octamer consists of two copies of two heterodimers -- histones H3/H4 and H2A/H2B [32]. Wrapping of DNA around histones limits the accessibility of the underlying DNA sequence to transcription factor binding and gene expression. Thus, the varying levels of chromatin condensation not only suggest a role for chromatin in the packaging of eukaryotic genomes, but more importantly in restricting or controlling gene expression[32].

In addition, the lysine-rich amino terminal tails of histones are subject to various chemical modifications such as lysine acetylation and methylation as well as serine phosphorylation at specific amino acid residues [33]. Some of these modifications have been detected near or directly over genes or transcribed regions and correlate with gene activity or silencing. Likewise, substitutions of the histone components of the octamer by variant proteins have been observed at specific sites in the genome and linked with gene activation [34]. Chemical modifications of histone tails and histone variant incorporation further underscore the role of chromatin structure in the control of gene expression in eukaryotic genomes. The persistence or inheritance of these chromatin modifications and structures associated with gene activation or silencing across cell division falls under the

study of "epigenetic" or non-sequence based mechanisms for transcription regulation [35,36].

1.2.2 Chromatin Immunoprecipitation with Microarrays

Chromatin immunoprecipitation with microarrays or ChIP-chip is among the tools that permit the study of genomic DNA in the context of chromatin in living cells. It is used to identify genomic DNA binding sites of transcription factors. It also allows the mapping of histones, histone variants, and specific histone tail modifications associated with genomic sequences. This strategy is an expansion of the chromatin immunoprecipitation (ChIP) method (Figure 1-2). Briefly, ChIP involves the chemical cross-linking of protein-DNA interactions in living cells by formaldehyde treatment. By sonication, genomic DNA is fragmented into lengths of ~1000bp, and then protein-DNA interactions of interest are isolated by immunoprecipitation (IP) with an antibody specific to the protein. The availability of a specific antibody is a key limiting step to defining the genomic DNA binding or association of a particular protein and has been circumvented by approaches using recombinant proteins [37,38]. Following enrichment for the protein-DNA complexes of interest, the chemical crosslinks are reversed. In the conventional method, the IP-enriched genomic DNA is assayed for specific fragments by Southern blot or polymerase chain reaction (PCR) to determine known or predicted protein association at specific genomic sequences [39].

The advent of microarrays has permitted the identification of IP-enriched genomic DNA fragments by hybridization of the DNA to probes tiled on microarray [40]. Thus, instead of testing a handful of sites for protein-DNA interaction, ChIP with microarrays

or ChIP-chip allows the simultaneous survey of thousands of sites if not more and represents a powerful tool for discovering novel sites of protein-DNA interaction. In this regard, the limitations of microarray use boil down to the coverage of the genome being surveyed and the resolution of that coverage. Initial use of ChIP-chip involved the tiling of known promoter regions from yeast as ~1000bp probes on a microarray[40,41]. Greater genome coverage and higher resolution involves the use of high-density microarrays with probes or oligonucleotide sequences less than 100bp in length representing genomic windows tiled with probes every 100bp or less [42-45].

The ability to achieve high-resolution and complete genome coverage for mammalian genomes not only required advances in microarray technology. The scale of the data set and the signal-to-noise ratio issues associated with high-throughout microarray data presented significant challenges for data analysis to be described in detail in Chapter 2. However, the push to perform genome-wide ChIP-chip in mammals was motivated by insightful studies in the budding yeast [46,47]. The yeast genome, less than 0.5% of the human genome in size, has been a productive ChIP-chip model system. Mapping of histones, histone modifications, histone variants, as well as the binding sites for nearly all known transcription factors in the yeast genome has provided a substantial preview of the lessons to be learned regarding gene regulation using ChIP-chip.

1.2.3 Nucleosome Depletion at Active Promoters

As described earlier, eukaryotic chromosomes consist of histone and non-histone chromosomal proteins along with DNA. The fundamental unit of chromatin organization defined by a nucleosome consists of a histone octamer around which 146bp of genomic

DNA is wound. Recent findings from genome-wide surveys of chromatin organization in yeast suggest that variation in nucleosome composition can be indicative of coding versus noncoding regions [45,48-50]. Although non-coding regulatory sequences have been identified by nuclease hypersensitivity on a gene-by-gene basis, the global identification of a generalized architecture of nucleosome depletion at promoters supports the model that these regulatory sequences must be accessible to transcription factors.

ChIP-chip experiments to map histone H3 and H4 consistently showed that intergenic regions are less densely occupied by nucleosomes than transcribed regions and that this depletion among intergenic regions is primarily associated with promoters upstream of transcribed sequences [49]. A recent pioneering strategy, combining micrococcal nuclease digestion to isolate mono-nucleosomes and DNA microarrays tiling 482kbp of the yeast genome at nearly every 20bp, resolved this stereotypical nucleosome-free region (NFR) to a roughly 150bp region situated about 200bp upstream of the start codon flanked on both sides by well-positioned nucleosomes [50]. NFRs were mainly associated with poly(dA-dT) stretches which have been shown to destabilize nucleosome formation *in vitro*. Rap1 consensus sites were also shown to be sufficient to induce NFRs and combinations of transcription factor motifs were strongly associated with these regions [48]. Whether sequence determinants or nucleosomal eviction by transcription factors cause these NFRs, this generalized promoter architecture in yeast clearly illustrates the functional role of chromatin organization in mediating transcription regulation.

1.2.4 Histone Modifications and Histone Variants at Promoters

Like nucleosome positioning, post-translational histone modifications also characteristically associate with distinct genomic loci. Two general classes of post-translational modifications, histone acetylation (H2A, H2B, H3, H4) and histone methylation (H3) have been mapped in genome-wide in yeast [45,51-53] and in a limited scale in higher eukaryotes [49,54-56]. The distinct mechanistic models for the addition of these modifications related to transcriptional activity might explain their characteristic bias for particular genomic loci.

The bias of histone acetylation sites near the beginning of genes is consistent with the model that transcriptional activators recruit histone acetyltransferases near regulatory regions such as promoters [45,51-53]. Mapping of individual histone acetyl-lysines has helped to partition the histone acetylation signals into transcription-dependent and transcription-independent modifications in yeast. In contrast with transcription-independent acetylation states, transcription-dependent lysine acetylation sites (H3K9, H3K14, H3K18, H4K5, H4K12, H2AK7) generally localize near 5' ends of transcriptionally active genes [52]. A ChIP-chip adaptation of the previously mentioned single nucleosome mapping strategy applied to these modification states in yeast precisely situates these transcription-dependent hyperacetylated histones at 5' sites flanking a hypoacetylated region consistent with the nucleosome-free promoter region[50,52].

On the other hand, histone H3K4 methylation in yeast has been clearly shown to be biased toward transcribed loci with the degree of methylation (mono-, di-, tri-) decreasing from the 5’ to the 3’ end [45,52]. This is consistent with the recruitment of the

H3K4 methyltransferase Set1 at the 5' end of regions actively transcribed by RNA Polymerase II (Pol II) [57]. Meanwhile, H3K36me3 has been observed throughout the coding regions consistent with the model that the matching histone methyltransferase Set2 is associated with the elongating form of Pol II [45].

To date, large scale histone acetylation and methylation studies in higher eukaryotes are generally consistent with findings in yeast, with some notable additions [49,54-56]. Mapping of H3K9/14 hyper-acetylation sites in activated human T cells correlate these sites highly with promoters. The acetylation sites, including those distal from known promoters, were also associated with previously identified conserved non-coding sequences based on human-mouse conservation and known regulatory elements in activated T-cells [55]. Mapping of H3K4me2 over orthologous human and mouse Hox clusters revealed broad regions of methylation over coding and intergenic regions which are notably not conserved in sequence but rather in location in human and mouse. Enrichment in intergenic transcription was detected over these methylated regions in the HoxA and HoxB loci in human and mouse, suggesting yet to be clarified mechanisms for the deposition of this modification state over coordinately regulated gene loci domains [56].

Like post-translational histone modifications, histone variant composition has also been implicated in transcriptional control. Recent results by ChIP-chip suggest that the histone H2A variant H2A.Z in yeast situates preferentially near promoter regions [58,59]. Single-nucleosome mapping resolved this H2A.Z enrichment over nucleosomes flanking the NFR [60]. The observation of H2A.Z at actively transcribed genes and inactive loci raises questions about when and how H2A.Z is deposited [58-61]. On the other hand, studies of the histone H3.3 variant in fly using biotin-tagged histone variants, in an adaptation of

the ChIP-chip procedure, show the pronounced enrichment of H3.3 over actively transcribed regions relative to canonical H3. This enrichment correlates with the level of transcription activity as well as Pol II occupancy and H3K4me2 levels derived from previous ChIP-chip studies [38,49]. A slight enrichment of H3.3 replication-independent replacement was also observed upstream of transcribed regions and slightly upstream of the region of nucleosomal depletion at promoters. Given a Pol II-associated model for H3.3 deposition over transcribed regions, it remains unclear how H3.3 is deposited upstream of promoters [38,61].

In summary, specific chromatin modifications and histone variant composition have been observed near promoters and correlated with gene activity (Figure 1-3). Although global observations were largely from studies in yeast, these patterns are predicted to be similar in higher eukaryotes. Furthermore, it is expected that distinct chromatin features also characterize other regulatory sites such as enhancers and silencers.

1.2.5 Sequence-Specific Transcription Factor Binding at Known Promoters

Ultimately, understanding the restricted expression of specific transcripts in a given cell type and condition requires the reconstruction of regulatory networks controlled by sequence-specific transcription factors. To this end, ChIP-chip using promoter arrays has been employed to define gene targets of several transcription factors implicated in development, disease, or cellular differentiation [62-65]. Usually in combination with mRNA expression profiling and motif-finding, these data sets have been used to predict groups of co-regulated genes activated or repressed by a factor or

combination of factors, as well as to define novel regulatory sequences for further verification. ChIP-chip for the binding of known transcription factors in yeast revealed regulatory network motifs such a single-input module, regulator cascade, multi-input module, auto-regulation, feed-forward loop, and multi-component loop[46,47]. Specifically these studies in yeast uncovered transcriptional modules – groups of transcription factors that share common target genes – such as those for the control of cell cycle progression and amino acid metabolism [47,66]. More recently, ChIP-chip studies of transcription factor binding at known promoters in human embryonic stem cells (hESC) uncovered a transcriptional module consisting of Nanog, Oct4, and Sox2 [63]. These factors bind together to at least 353 gene promoters in hESC and form a feed-forward loop, as well as an interconnected auto-regulatory loop. Among their target genes are key transcriptional regulators of cell proliferation and differentiation, demonstrating the central role of this transcriptional module in the maintenance of hESC self-renewal and pluripotency.

1.3 Overview of the Dissertation

In the preceding sections I highlighted our limited understanding of the control of mammalian gene expression, in particular of the regulatory DNA elements such as promoters and enhancers which mediate the controlled expression of subsets of genes in specific cell types and tissues. I also described the pioneering use of ChIP-chip for large-scale mapping of protein-DNA interactions in yeast and to a limited extent in flies and mammals. These maps underscore the power of ChIP-chip in characterizing regulatory DNA elements such as promoters by distinct chromatin features as well as by the binding of sequence-specific transcription factors. Thus, the work described in this dissertation

involves a pilot study to test the feasibility of genome-wide ChIP-chip in a mammalian cell type in order to map and profile active promoters at high-resolution and in an unbiased fashion (Chapter 3). The second major section involves the application of this genome-wide approach to identify active promoters in a panel of mammalian organs and embryonic stem cells in order to characterize the promoter sequence and epigenetic features of tissue-specific expression (Chapter 4). These two chapters in full represent published work and a manuscript in preparation, respectively.

Aside from providing global views of promoter activity, a common thread underlying those two major projects is the development of analysis strategies for the new kind of ChIP-chip data generated. In Chapter 2, I provide an overview of ChIP-chip data analysis issues and the common methods I have adapted to characterize six high-resolution genome-scale ChIP-chip studies in human and mouse generated for these projects. In particular, I highlight a model-based approach for defining protein-DNA interactions from ChIP-chip data (2.3).

Acknowledgments

Chapter 1, in part quotes sections from: Barrera, Leah O.; Ren, Bing. The transcriptional regulatory code of eukaryotic cells, Current Opinion in Cell Biology, Vol. 18, 2006. I was a primary author of this review.

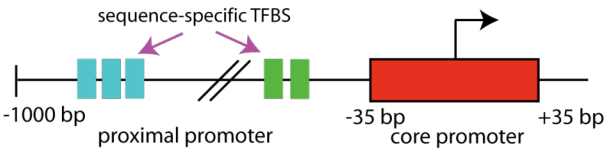


Figure 1-1. Promoter schematic.

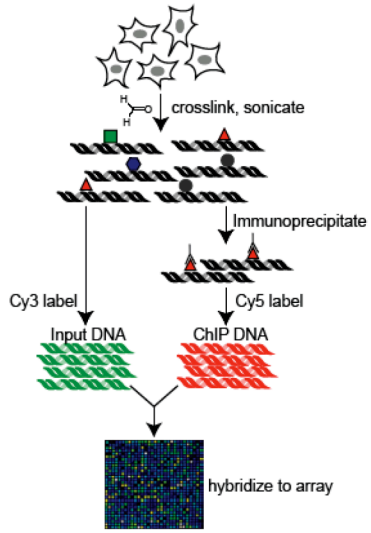


Figure 1-2. ChIP-chip strategy

(From publication: Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press.)

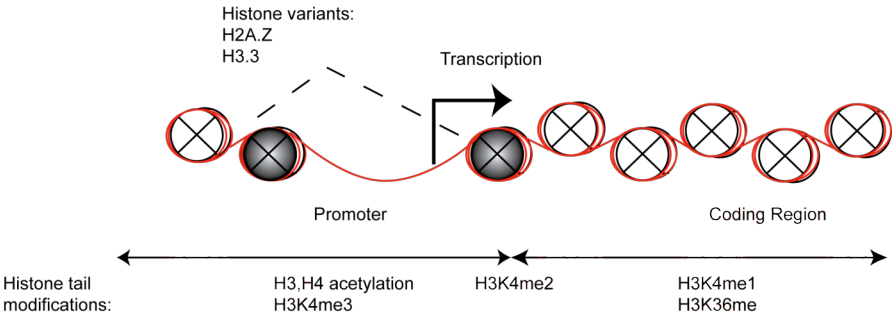


Figure 1-3. Chromatin structure and modification at active promoters

General view based on ChIP-chip studies from model organisms (yeast and fly).

Chapter 2

ChIP-chip Data Analysis

The advent of ChIP-chip using genomic tiling arrays has led to a growing number of analysis methods for pre-processing, binding site identification, and high-level analysis [67-72]. Genomic tiling arrays contain oligonucleotide or PCR products which cover the genome in a tiling path. Compared with low resolution (~1000bp) PCR arrays, high-resolution (~100bp) oligonucleotide tiling arrays reveal ChIP-chip binding events by the enrichment of multiple neighboring probes instead of a single probe, thus permitting more reliable binding site identification. Advances in high-density oligonucleotide array technology have enabled the genome-wide coverage of non-repetitive sequences in mammalian organisms such as human and mouse. Experiments using these arrays typically result in a large volume of data, posing challenges in data analysis. In this chapter, we review the strategies we have developed and adapted for the analysis of the resulting data used for the pilot applications described in Chapter 3 and 4. Given the growth and widespread adoption of the technology, we also highlight notable ChIP-chip analysis approaches concomitant with and after our work.

2.1 Analysis Overview

Steps involved in the analysis of ChIP-chip data are largely analogous to the steps involved in the analysis of microarray gene expression data (Figure 2-1). Some common steps are data pre-processing, annotation and high-level analysis, database management and submission to microarray repositories. Analogous steps include binding site

identification for ChIP-chip, compared with expression or 'Present' calls in microarray gene expression. High-level analyses for ChIP-chip data involves the adaptation of visualization strategies to examine patterns of genomic binding and the use of motif-finding strategies to find a common motif (or motifs) characterizing the binding sites for a particular factor. Given the extensive literature for microarray expression data analysis, we only highlight key adaptations and developments for normalization, binding site identification and higher-level analysis[73]. Implementation of the following analyses requires standard bioinformatics capabilities – Perl scripting, database management systems such as MySQL, statistics modules from R, motif-finding programs, and specialized software for binding site identification.

2.1.1 Microarray Platforms

Coverage and resolution

Prior to the advent of genomic tiling arrays, mammalian ChIP-chip arrays typically had biased coverage and limited resolution [42,44,64,74-78]. Coverage refers to the genomic regions surveyed by the probes on the array, while tiling resolution refers to the average distance (center to center) of the genomic positions corresponding to any two adjacent probes tiling a genomic region (Figure 2-2). In early mammalian ChIP-chip applications, array coverage was biased to known promoters or CpG islands with one probe or array feature corresponding to each promoter or CpG island [64,79,80]. Array designs for unbiased coverage of genomic regions initially covered selected chromosomes or genomic regions in a tiling path. Pilot applications include surveys of the smallest (autosomal) chromosomes in the human genome, chromosomes 21 and 22,

as well as 44 loci, comprising 1% of the human genome, designated as Encyclopedia of DNA Elements (ENCODE) regions [42,74,75,81]. Mammalian genome-wide high-resolution tiling arrays using the NimbleGen array platform were first described in a publication corresponding to Chapter 3 [43]. Coverage of the entire non-repetitive human genomic sequence at 100 base pair (bp) resolution required 38 arrays. Elimination of repetitive sequences is required for unambiguous mapping of ChIP-enrichment signal to genomic positions.

Platforms

The specifications and analysis support for the three commercial array platforms most commonly used and currently available for ChIP-chip with genomic tiling arrays – Affymetrix, Agilent, and NimbleGen – are summarized in Figure 2-3. These specifications are compared to a prototypical PCR array made "in-house" by individual laboratories [81]. Currently, all platforms offer genome-wide array coverage of mammalian genomes such as mouse and human. To date, the Affymetrix array platform offers the highest density and tiling resolution, covering the entire human genome at 35 bp resolution with 7 arrays. On the other hand, the NimbleGen and Agilent array platforms are based on more flexible array synthesis technologies which do not require physical masks for every array design. In particular the mask-less array synthesis (MAS) technology underlying NimbleGen arrays permits custom array designs for virtually any sequenced genome. We refer the interested reader to the following references describing the array production technology underlying each platform [82,83].

2.1.2 Data Pre-Processing

After array scanning and image extraction, each probe or array feature is associated with an intensity signal or a pair of intensity signals corresponding to the amount of labeled DNA bound. In addition to image quantification, pre-processing of this ChIP-chip microarray data typically requires data normalization. There are two main types of normalization: (1) within-array normalization, and (2) between-array normalization for replicates and sample comparison. To the best of our knowledge, there are no extensive reviews of normalization methods for ChIP-chip data compared with gene expression data. However, as with gene expression data, the goal in normalizing ChIP-chip data is to identify and remove systematic sources of variation in the measured intensities in order to better resolve ChIP enrichment based on biological variation. For two-color data, ChIP enrichment measured by a probe is typically represented as the log of the ratio of the intensity of the bound Cy5 labeled DNA or *R* (ChIP enriched) over the intensity of the bound Cy3 labeled DNA or *G* (input DNA). Most normalization schemes simultaneously take the intensities for each channel as input and output a normalized log ratio [73,84].

Intensity-dependent normalization

We use an R normalization package originally developed for microarray gene expression data, *limma*, to normalize ChIP-chip tiling array data [85]. We generally outline the procedure we use for ChIP log ratio normalization of two-color data (NimbleGen platform), but recommend the *limma* user's guide for more detailed description of specific functions.

For two-color data, we extract the two channel intensities and an identifier for each probe and load them into the *RGList* object, a data structure, used as input to the various normalization functions in *limma*. To assess the quality of the data before and after normalization, we view the traditional M versus A plots. M refers to the log of the ChIP enrichment ratio, $\log_2 \frac{R}{G}$ and A refers to the average log intensity from both channels, $\frac{1}{2}(\log_2 R + \log_2 G)$, for each probe. Thus, the M vs. A plot is a scatter plot where the log ratio for each probe is plotted relative to its average log signal. In general, features on the plot are expected to be distributed near the M=0 line, with true ChIP enrichment represented by outliers at high values of M. Anomalous intensity-dependent effects in which high M associates with low A can be spotted visually using these spots [84] (Figure 2-4).

Loess is a non-linear normalization strategy successfully used for the intensity-dependent normalization of two-color expression array data [84]. It is a scatter-plot smoother that performs robust locally linear fits[86]. This then results in the shifting of M values up or down depending on their intensity A according to the loess fit *c*:

$$\log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c(A_i)$$

We generally use loess for within-array normalization to adjust ChIP log ratio values and specify only the control probes on the array to calculate the loess fit (*normalizeWithinArrays*). The use of only control probes for normalization is critical when a large proportion of arrayed features map to binding sites for a factor (Figure 2-4).

We typically apply this normalization strategy when analyzing ChIP-chip data without replicates.

Dye-swap normalization

Dye incorporation or signal intensity biases can be treated by a simple dye-swap normalization [87]. This requires an experimental design in which the ChIP-enriched and genomic input samples for one of a pair of technical replicates for a ChIP-chip experiment are reversely labeled using Cy3 and Cy5, respectively, instead of the typical assignment of Cy5 to ChIP-enriched DNA and Cy3 to the genomic input DNA. The ChIP log ratio enrichment for each probe from the pair of dye-swapped technical replicates is obtained by simply taking the average of the log ratio from each replicate where the ratio is based on the intensity signal from the ChIP-enriched channel over the genomic input channel. We applied this type of normalization for replicate arrays used for condensed ChIP-chip scans in Chapter 3.

Sequence-based normalization

As an alternative to the normalization methods in *limma* or dye-swap normalization, Model-based Analysis of 2-Color arrays (*MA2C*) was recently developed to adjust probe intensities based on GC-content (XS Liu, unpublished). MA2C extends the Model-based Analysis of Tiling-array (MAT) method recommended for normalizing single-channel ChIP-chip data based on the Affymetrix tiling array platform [70]. MAT takes as input the raw CEL and BPMAP files and adjusts the single-channel intensity for each 25 bp probe, *i*, based using the following model of probe behavior based on probe sequence content:

$$\log(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k \in \{A,C,G\}} \beta_{jk} I_{ijk} + \sum_{k \in \{A,C,G,T\}} \gamma_k n_{ik}^2 + \delta \log(c_i) + \varepsilon_i$$

Here the intercept α is based on the "T" count, β is the position effect of each nucleotide, γ is the effect of the nucleotide count squared, δ the effect of probe copy number c_i, and ε is the error term for the probe *i*. The value n_{ik} is the nucleotide *k* count in probe *i*, and I_{ijk} is the indicator function such that I_{ijk} is 1 if the nucleotide at position *j* is *k* for probe *i.* and 0 otherwise.

The parameters for the model are estimated with standard least-squares regression using all probes on the array. Probes in the array are then grouped based on sequence content similarity into "affinity bins" of ~ 3000 probes. Finally, each probe signal is standardized relative to its expected probe behavior and the standard deviation within its affinity bin:

$$t_i = \frac{\log(PM_i) - \hat{m}_i}{s_{i,affinitybin}}$$

This correction significantly improves ChIP-chip analysis of Affymetrix data, and suggests the importance of modeling baseline probe behavior for short oligonucleotides (25 bp) used in this platform[70]. In fact, a model-based correction of probe intensities based on sequence content has also been successful for the adjustment of Affymetrix microarray gene expression data using the R package *gcrma* [88].

Quantile normalization

To normalize across a set of replicate arrays we apply quantile normalization. This method forces the distribution of array probe values in each array across a set of arrays to be the same [89]. In other words, the value at the 92nd quantile for array A should

be equal to the value at the 92nd quantile for array B after quantile normalization. We use the *normalizeBetweenArrays* function in *limma*. Quantile normalization strategies are also available through other R microarray analysis packages *affy* and *rma.*

2.1.3 Single Array Error Model

In this section, we briefly review the Single-Array Error Model, which is the standard for the ChIP-chip binding site analysis of low-resolution PCR arrays[40,46,47,64,90]. It is a one-sided test for ChIP enrichment calculated for each probe. The single array error model is based on the simple calculation of a test statistic for each probe testing the null hypothesis that there is no difference in intensities for the Cy5 (*R*) and Cy3 (*G*) channels, $R-G=0$, against the one-sided alternative that the difference is greater than 0, $R-G>0$.

For each array feature, summary statistics are given in the GPR file for the intensity values measured in each channel (F635 Median, F635 Mean, F532 Median, F532 Mean, F635 SD, F532 SD, F635 +1SD, F532 +1SD, B635 Median, B532 Median, Flags). We associate each array feature with a Cy5 (F635 channel) and a Cy3 (F532 channel) intensity value by taking the median statistics for the foreground measurements and subtracting (non-negative) median statistics for the background measurements:

$$Cy5 = F635Median - B635Median$$

$$Cy3 = F532Median - B532Median$$

If there are blank spots on the array adjust the Cy5 and Cy3 intensities further by subtracting the median intensity measured at each channel among the blank spots in the array.

To normalize the Cy3 intensities relative to the Cy5 intensities, the median of the intensity ratios (Cy5/Cy3) is used to adjust each of the Cy3 intensities in a global scaling normalization:

$$Cy3_{normalized} = Cy3 \times Median\left(\frac{Cy5}{Cy3}\right)$$

A scatter-plot of log(Cy5) relative to log(Cy3) intensities is used to qualitatively assess the ChIP-chip experiment. A relatively tight distribution of spots over the 45-degree line suggests a clean hybridization. Deviations from the 45-degree line with higher Cy5 relative to Cy3 intensities suggest array features overlapping ChIP-enriched genomic loci (Assumption: ChIP-enriched DNA is labeled using the Cy5 dye). (Figure 2-5)
The Single-Array Error Model (SAEM) test statistic for each array feature is then calculated as:

$$X = \frac{(Cy5 - Cy3)}{\sqrt{[\sigma_{Cy5}^2 + \sigma_{Cy3}^2 + f^2(Cy5^2 + Cy3^2)]}}$$

The $\sigma_{Cy5}, \sigma_{Cy3}$ values are the F635 SD and F532 SD measurements for each array feature, respectively. The function f is a fractional multiplicative error due to hybridization non-uniformities, fluctuations in dye incorporation efficiency, and scanner gain fluctuations [90]. f is chosen so that X has unit variance. It is estimated once from control experiments in which Cy5 and Cy3 labeled DNA come from the same reference sample [90]
A one-sided *p*-value for the standardized SAEM test statistic (*X*-statistic) is calculated using the standard normal cumulative distribution function (cdf) $N(0,1)$ or Φ:

$$p = 1 - \Phi\left[\frac{X - \mu}{\sigma}\right]$$

This requires the assumption that the *X*-statistics have a Gaussian distribution with a minor right-tail corresponding to array features overlapping ChIP-enriched genomic loci. When the normality assumption does not hold as observed when the distribution is heavily skewed, alternative methods for modeling the null distribution of the *X*-statistics can be used to assign *p*-values or False-Discovery Rate (FDR) adjusted significance values [91]. The use of FDR-adjusted *p*-values obviates concerns regarding simultaneous or multiple testing of tens of thousands of probes. Array features overlapping ChIP enriched genomic loci are then selected based on a particular *p*-value or FDR threshold (for example $p < 0.001$ or 5% FDR) and classified as "binding sites".

2.2 Binding Site Identification

The key goal in the majority of ChIP-chip experiments is to define DNA regions bound by the factor of interest. For ChIP-chip using microarrays spotted with PCR fragments an adaptation of the single-array error model (SAE) is commonly used to define arrayed features as overlapping ChIP-enriched genomic DNA and was described in the last section. Binding site identification for ChIP-chip using high-resolution oligonucleotide tiling arrays requires integrating intensity information over neighboring probes to determine enrichment. In this case, it is not sufficient for a single array feature to have enriched signal, but rather enrichment is determined over a genomic window of probes (Figure 2-6).

For initial screens at 100 bp resolution, we have used a simple windowing approach to define ChIP-enriched genomic regions as being spanned by a minimum of 4 probes separated by a maximum of 500bp with ChIP enrichment log ratios greater than 2.5 standard deviations from the mean. This min-run, max-gap approach has been used for the analysis of tiling array data for transcriptome profiling [92]. An analogous windowing approach called ChIPOTle (ChIP On Tiled arrays) reassigns the value for each probe based on the average of log ratios for all probes within a pre-specified genomic window size (for instance, 1000 bp) centered on that probe [68].

These approaches, however, do not adequately take advantage of the profile of ChIP enrichment. Recently developed strategies to model genomic neighborhood enrichment include Joint Binding Deconvolution (JBD) and TILEHGMM [69,71]. JBD, in particular, requires additional information in the form of the fragment length distribution of the sonicated DNA for each ChIP-chip experiment to predict binding sites [71]. To precisely predict binding sites for a given factor at probe-level resolution, we have effectively used a peakfinding strategy called MPeak [72]. This method models ChIP enrichment at a transcription factor binding site to be shaped like peaks or triangles over a genomic window, given assumptions regarding the sonication and ChIP steps (Figure 2-7). It uses a fast algorithm for determining non-overlapping binding peaks for a user-selected significance level. We describe the derivation and implementation of this approach in detail in the following section.

2.3 Peakfinding Model

Our approach to peak-finding is model-based because we derive a probability model for the ChIP-chip process which explains the resulting ChIP-chip data as outlined below:

1. Genomic binding sites: Factor binding sites (such as promoters) on the genome can be idealized as a set of points on the real line. Let's denote the locations of these binding sites by their coordinates $B_1, B_2, \ldots, B_M$. The total number M of binding sites and their coordinates are unknown, and they need to be inferred from the ChIP-chip data.
2. Protein binding: In the ChIP-chip experiment, the proteins are bound to their cognate binding sites. For a genome sequence, let p_m be the probability that the binding site m is bound by a protein. Different binding sites are assumed to be independent of each other.
3. Sonication: The sonication process fragments chromosomes into shorter DNA fragments (~1000 bp). Each fragment is an interval on the real line. For a genome sequence, the set of cut points are randomly distributed. A simple probability model is the Poisson point process, which has the following assumptions: (1) the probability that a cut point occurs in a small interval $(x, x+\Delta x)$ is $\lambda(x)\Delta x$, where $\lambda(x)$ is the intensity function measuring how dense the cut points are around x. $1/\lambda(x)$ can be considered the expected length of the intervals between two consecutive cut points around x. (2) For non-overlapping intervals, what is happening in one interval is independent of what is happening in the other interval.

4. Immunoprecipitation: For each protein bound to a binding site, the probability that it is recognized and bound by the antibody is α. For a DNA fragment to be immuno-precipitated, it must contain at least one binding site that is bound by the protein, which must in turn be recognized and bound by the antibody. We call such a binding site a "good binding site." Thus, the probability that B_M is a good binding site is $p_m\alpha = q_m$. A DNA fragment that contains at least one good binding site is called a "good fragment."
5. Tiling array of probes: At each location x, the array signal measured by a probe at x is denoted by $Y(x) = log(Cy5/Cy3)$. It measures the relative abundance of good fragments that contain x. The actual binding sites are generally several bp long (not a point-source), and a probe can be as long as 50 bp. Here we mathematically idealize them as dimension-less points on the real line for simplicity.

2.3.1 Probability Model

Consider a random genome sequence. The ChIP process produces from this genome sequence a collection of non-overlapping good fragments. These good fragments only cover part of the whole genome. For any location x, let $p(x)$ be the probability that x is covered by a good fragment. In the experiment, there are a large number of genome sequences, and $p(x)$ manifests itself as the concentration of good fragments covering x. So $p(x)$ can be considered the theoretical prediction of the signal value measured by probe x. In the following, we shall calculate $p(x)$ under various scenarios. In order to make this subsection easy to follow, we add some non-rigorous steps in the derivations.

A key observation is that for x to be covered by a good fragment, a necessary and sufficient condition is that there is no cut point between x and at least one good binding site.

One binding site scenario: Let's first consider the simplest scenario where there is only one binding site at the origin of the real line. Then:

$$\begin{aligned} p(x) &= \Pr(0 \text{ is a good binding site and no cut point exists between 0 and } x) \\ &= q \times \Pr(\text{no cut} \in (0, x)) \end{aligned}$$

where q is the probability that 0 is a good binding site, i.e., it is bound by a protein, which is in turn bound by the antibody. Without loss of generality, let's assume that $x > 0$.

To compute $\Pr(\text{no cut} \in (0, x))$, we can divide the interval (0, x) into a large number of small bins, $(0, \Delta x), (\Delta x, 2\Delta x), ..., (i\Delta x, (i+1)\Delta x), ..., ((n-1)\Delta x, n\Delta x)$ where $\Delta x = x / n$. Let $x_i = x / n$. According to the Poisson assumption,

$$\begin{aligned} \log \Pr(\text{no cut} \in (0, x)) &= \sum_{i=1}^{n} \log(1 - \lambda(x_i)\Delta x) \\ &\to -\int_0^x \lambda(s) ds, \text{ as } n \to \infty \end{aligned} \qquad \text{(Equation 1)}$$

The last step follows the Taylor expansion: $\log(1 - \lambda(x_i)\Delta x) = -\lambda(x_i)\Delta x + o(\Delta x)$, with $o(\Delta x)$ being a term that decreases to 0 faster than $1/n$ as $n \to \infty$. Thus

$$\log p(x) = \log q - \int_0^x \lambda(s) ds, \text{ for } x > 0.$$

If we assume $\lambda(x) = a$ for $x > 0$, then $\log p(x) = c - ax$, for $x > 0$, where $c = \log q$.

Similarly for $x \le 0$, if we assume $\lambda(x) = b$, then $\log p(x) = c + bx$, for $x \le 0$. We can combine the two equations for $x > 0$ and $x \le 0$ into one equation,

$$\log p(x) = c - b[-x]^+ - a[x]^+ \quad \text{(Equation 2)}$$

where $[x]^+ = x$ if $x > 0$, and $[x]^+ = 0$ otherwise.

Equation (2) has a triangle shape peaked at 0, and is the basis for our model-based peak recognition method. However, this model assumes that there is only one binding site. For real data, the above model is true only around a local neighborhood of a binding site, where the effects from other binding sites can be neglected. In the following, we shall study the situation where there are more than one binding sites, in order to understand how different binding sites affect each other.

Two binding site scenario: Suppose there are two binding sites B_1 and B_2. Let's assume that $B_1 \le B_2$. Let q_1 and q_2 be the probabilities that they are good binding sites, respectively. For $x \in (B_1, B_2)$, $p(x)$ is influenced by both B_1 and B_2.

$$\begin{aligned} p(x) &= \Pr(B_1 \text{ is good and no cut} \in (B_1, x) \text{ or } B_2 \text{ is good and no cut} \in (x, B_2)) \\ &= q_1 \exp\{-\int_{B_1}^{x} \lambda(s)ds\} + q_2 \exp\{-\int_{x}^{B_2} \lambda(s)ds\} - q_1 q_2 \exp\{-\int_{B_1}^{B_2} \lambda(s)ds\} \end{aligned} \quad \text{(Equation 3)}$$

where the last step follows the same logic as Equation (1).

If B_1 and B_2 are far away from each other, and if x is close to B_1, then the last two terms in Equation (3) can be neglected, and we will obtain an approximated equation that is in the same form as (2) in the one binding site scenario.

General scenario: Now we are ready to derive the formula for general scenario, where there are M binding sites $B_1, \ldots, B_M$. For notational convenience, we also add $B_0 = -\infty$, and $B_{M+1} = \infty$, with $q_0 = q_{M+1} = 0$. For $x \in (B_M, B_{M+1})$,

$$p(x) = \Pr(\text{no cut} \in (x, \text{nearest good binding site to the left})$$
$$\text{or no cut} \in (x, \text{nearest good binding site to the right})) \qquad \text{(Equation 4)}$$
$$= p_L(x) + p_R(x) - p_L(x)p_R(x)$$

where

$$p_L(x) = \Pr(\text{no cut} \in (x, \text{nearest good binding site to the left}))$$
$$= \sum_{i=0}^{m} \Pr(\text{nearest good binding site to the left is } B_i \text{ and no cut} \in (B_i, x)) \quad \text{(Equation 5)}$$
$$= \sum_{i=0}^{m} \left[\prod_{j=i+1}^{m} (1 - q_i) \right] q_i \exp\{-\int_x^{B_i} \lambda(s)ds\}.$$

$$p_R(x) = \Pr(\text{no cut} \in (x, \text{nearest good binding site to the right}))$$
$$= \sum_{i=m+1}^{M+1} \left[\prod_{j=m+1}^{i=1} (1 - q_i) \right] q_i \exp\{-\int_{B_i}^{x} \lambda(s)ds\}. \qquad \text{(Equation 6)}$$

Using equations (5) and (6), $p(x)$ can be calculated according to Equation (4). From the above analysis, we can see that the triangle shape fits the data only within a local range around a true binding site. So in the data analysis, we shall fit a truncated triangle shape model whose range is adaptively determined.

Chip measurement

The "chip" step of the ChIP-chip process measures log *p(x)*. The Cy5 measures the abundance of DNA fragments in the IP-enriched DNA pool, and Cy3 measures the abundance of DNA fragments in the un-enriched DNA pool. For a DNA fragment containing probe *x*, the hybridization strength, i.e., the probability that it will be hybridized by the probe *x*, can depend on *x*. By calculating $Y(x) = \log(Cy5 / Cy3)$, this dependence is cancelled out. We shall simply assume that the observational errors are additive and follow a stationary Gaussian process.

2.3.2 Peak Identification

The previous section shows that a binding site causes an approximately truncated triangle shape for the signals of the probes around this binding site. In this sub-section, we propose a model-based method to recognize these shapes. After finding these truncated triangle shapes, including their positions and ranges, we can pool the probe signals within the range of each identified shape to test against the background noise hypothesis, to decide whether these signals are caused by a true binding site.

Truncated triangle shape model

We fit the truncated triangle shape model to the data around each probe in order to identify the positions and ranges of the shapes.

Let's use x_0 to denote the genomic coordinate of this probe. We look at a window around x_0. Let L be the number of probes to the left of x_0 within the window. Let R be the number of probes to the right of x_0 within the window. Let's denote the genomic coordinates of the probes to the left of x_0 by $(x_{-L},...x_{-1})$, and the coordinates of the probes to the right of x_0 by $(x_1,...x_R)$. Let the signals measured by these probes be $(y_{-L},...y_{-1},y_0,...y_R)$. We then fit the following multiple regression model:

$$y_i = c - b[x_0 - x_i]^+ - a[x_i - x_0]^+ + \varepsilon_i, -L \le i \le R, \quad \text{(Equation 7)}$$

where $a \ge 0$ and $b \ge 0$. We fit this model by a constrained least squares method. Let $Y = (y_i)_{i=-L}^{R}$ and $X = (1, -[x_0 - x_i]^+, -[x_i - x_0]^+)_{i=-L}^{R}$. Then, the least squares estimates of the coefficients are $(\hat{c}, \tilde{b}, \tilde{a})' = (X'X)^{-1}X'Y$. To satisfy the positivity constraints, we let $\hat{a} = [\tilde{a}]^+$ and $\hat{b} = [\tilde{b}]^+$. Because of DNA packaging and interactions with histones etc.,

there is reason to believe that the chopping rates around different binding sites may be different during the sonication step. Therefore we assume that each peak has its own slopes, *a* and *b*.

Let $\hat{Y} = X(\hat{c}, \hat{b}, \hat{a})$. We calculate the residual variance $\hat{\sigma}^2 = \left\| Y - \hat{Y} \right\|^2 \Big/ (L + R + 1 - d)$ where *d* is the number of regression coefficients. If both *L* and *R* are non-zero, then $d = 3$. If $L = 0$ or $R = 0$ then $d = 2$.

The residual variance $\hat{\sigma}^2$ is used for identifying the peak positions as well as the ranges *L* and *R*. It is not used for testing the significance of the peaks. Specifically, model (7) is correct under the following two assumptions:

1) x_0 is a true binding site, and

2) *λ(s)* is constant within $[-L, 0)$ and $(0, \mathrm{R}]$, respectively.

If neither assumption is correct, then model (7) is incorrect, and the residual variance $\hat{\sigma}^2$ will include the contribution from model bias. Therefore, a true binding site can be detected by the local minimum of the fitted $\hat{\sigma}^2$.

To be more specific, for any x_0 and *L,R,* let the signal $y_i = f(x_i) + \varepsilon_i$. $f(x)$ is a truncated triangle shape peaked at x_0 if and only if assumptions 1) and 2) hold. If x_0 is not a true binding site, then $f(x)$ will not be a truncated triangle shape peaked at x_0. Instead, it will be a triangle peaked at a binding site other than x_0. Let $f = \left(f(x_i)\right)_{i=-L}^{R}$ and $\varepsilon = (\varepsilon_i)_{i=-L}^{R}$. We can write $Y = f + \varepsilon$. Let $H = X(X'X)^{-1}X'$ be the projection matrix, and let $\hat{Y} = HY, \hat{f} = Hf$, and $\hat{\varepsilon} = H\varepsilon$ be respectively, the projections of

$Y, f,$ and ε onto the space spanned by X. Then $E\left\|Y-\hat{Y}\right\|^2 = \left\|f-\hat{f}\right\|^2 + E\left\|\varepsilon-\hat{\varepsilon}\right\|^2$ because $E[\varepsilon]=0$. If assumptions 1) and 2) hold, then $f(x_i) = c - b[x_0 - x_i]_+ - a[x_i - x_0]_+$, so $\left\|f-\hat{f}\right\|^2 = 0.$ If we shift x_0 from the true binding site while keeping *L* and *R* fixed, then $\left\|f-\hat{f}\right\|^2 > 0.$ Assuming ε_i come from a stationary process, and assuming that the probes are equally spaced, then $E\left\|\varepsilon-\hat{\varepsilon}\right\|^2$ remains unchanged under the shift, because *X* remains the same. Therefore, $E\left\|Y-\hat{Y}\right\|^2$ or $E(\hat{\sigma}^2)$ is a local minimum relative to the shifting operation if assumptions 1) and 2) hold. This fact does not depend on the assumption that ε_i are uncorrelated. Therefore, we may use the residual variance $\hat{\sigma}^2$ to identify the locations of the binding sites.

We also use the residual variance $\hat{\sigma}^2$ to determine the ranges *L* and *R* of the truncated triangle shape. If ε_i are uncorrelated with constant marginal variance $\hat{\sigma}^2$, then under assumption 1), $E(\hat{\sigma}^2) = \sigma^2$ for any *L* and *R* that satisfy assumption 2). If *L* or *R* is too large for assumption 2) to be true because of the effects from nearby binding sites, then $E(\hat{\sigma}^2) > \sigma^2$. In practice, we choose *L* and *R* that give us minimum $\hat{\sigma}^2$ among all the allowable combinations of *L* and *R*. This is a conservative choice. *L* and *R* determine the range of a fitted triangle shape, so that we can pool the signals within this range, and use their average to test against the background hypothesis.

For a peak shape caused by a true binding site, the conservative choice of *L* and *R* already enables us to include the strong signals around the binding site. Even though the conservative choice of *L* and *R* may fail to include the relatively weak signals of the

probes that are near the two ends of the true triangle shape, we will not lose much power in testing against the background hypothesis. At the same time, if x_0 is not a true binding site, then such choice of *L* and *R* will prevent us from pooling signals that may be caused by nearby binding sites, so that we will not declare too many false positives.

If ε_i are stationary but not uncorrelated, with marginal variance $\hat{\sigma}^2$, then under assumptions 1) and 2), $E\|\varepsilon - \hat{\varepsilon}\|^2 = E\|\varepsilon\|^2 - E\|\hat{\varepsilon}\|^2 = (L + R + 1 - tr(H\Sigma))\sigma^2$, where $\Sigma = E(\varepsilon\varepsilon')/\sigma^2$ is the correlation matrix of ε.

$E(\hat{\sigma}^2) = \sigma^2(L + R + 1 = tr(H\Sigma))/(L + R + 1 - d)$ which depends on *L* and *R* and is not an unbiased estimate of the marginal variance $\hat{\sigma}^2$. In this situation, we continue to choose *L* and *R* with minimum $\hat{\sigma}^2$.

Sometimes, ChIP-chip may produce an enriched region as a plateau of high values instead of a peak. In this case, our method can still detect a peak from such a region because the truncated triangle shape model can fit such plateau shape with very flat slopes. Occasionally, some probes may fail to function normally during the ChIP-chip experiment. Such dysfunctional probes may produce overly small or large signals. The truncated triangle shape model enables us to detect and remove such probes as outliers.

2.3.3 Peakfinding algorithm

1. Identify all the local maximum probes in the data. A probe is a local maximum probe if its signal is greater than all the signals within *k* bp away (*k* is a parameter that is pre-specified and the default value is 200).

2. As a starting point, pick the probe with the largest signal among all the local maximum probes.
3. At the current probe *x*, fit the triangle shape model as described above, for all combinations of *(L,R),* where both *L* and *R* are chosen within a range from the smallest allowable value to the largest allowable value (these two values are pre-specified, and the default numbers are 300 bp and 1500 bp respectively). Then choose the *(L,R)* that gives the smallest residual variance $\hat{\sigma}^2$. We call *(x − L, x + R)* the range of this probe *x*, and $\hat{\sigma}^2$ the residual variance of *x*.
4. Repeat the above model fitting procedure for the neighbors of this current local maximum probe. For each neighboring probe *x*, obtain its range and residual variance as described in step 2. Then among the current local maximum probe and its neighbors, choose the probe with the smallest residual variance to identify the best fitted triangle shape. We mark this probe as a potential binding site.
5. For any local maximum probe other than the above marked probe within the range of this best fitted triangle shape, we compare the fitted value of the best fitted triangle and the fitted value of the triangle centered at this local maximum probe. If the difference between the two fitted values at this local maximum probe is less than a threshold (which is a factor times the standard deviation of the residuals of the best fitted triangle, e.g default factor is 1.5), then this local maximum probe is said to be explained by the best fitted triangle and it is marked as non-peak.

6. Among all the local maximum probes still not marked, choose the local maximum probe with the largest signal. Then, go back to step 3. Stop the algorithm if all the local maxima are marked.

2.3.4 Evaluating Peak Significance

For a potential binding site *x*, suppose the truncated triangle shape fitted at *x* covers *n* probes. Let $Y_1, Y_2, ..., Y_n$ be the signals of these *n* probes, which can be considered the signals caused by the potential binding site *x*. We want to test whether *x* is a real binding site by pooling these *n* probes. We decide to use the following test statistic: $\overline{Y}_n = \sum_{i=1}^{n} Y_i / \sqrt{n}$. A similar method was proposed by Buck, Nobel and Lieb to calculate the significance of the sliding window average for each probe[68].

If $Y_1, Y_2, ..., Y_n$ are not caused by a binding site, they should be pure noise, which can be modeled by a stationary process. This process, however, is not independent white noise, because there are auto-correlations between nearby probes. We may assume that Y_i is correlated with its neighbors Y_j with $|P_j - P_i| \le m$ (P_j and P_i are the genomic positions of Y_j and Y_i, respectively). Then

$$\begin{aligned} Var(\overline{Y}_n) &= Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i,j} Cov(Y_i, Y_j) = \frac{1}{n}\sum_{|P_i - P_j| \le m} Cov(Y_i, Y_j) \\ &\approx Var(Y_i)\left(1 + \sum_{|P_i - P_j| \le m} Cov(Y_i, Y_j)/Var(Y_i)\right) = \gamma^2(1+f) \end{aligned} \quad \text{(Equation 8)}$$

where γ^2 is the marginal variance $Var(Y_i)$, and *f* is the auto-correlation factor. Both can be estimated from the data. Specifically, we can first calculate the marginal standard deviation of the whole sequence of signals. Then we remove those signals that are above

a threshold (default value is 2.5 times the marginal standard deviation). After that we estimate γ^2 and f based on the remaining signals. Because the true peak shapes only occupy a few parts of the whole sequence, and the vast majority of the signals are background noise, such a procedure gives reasonable estimates of γ^2 and f. We calculate the *p*-value by comparing the observed $\overline{Y}_n$ with $N(0, \gamma^2(1+f))$. The normal distribution can be justified by the central limit theorem. We can trim the insignificant peak shapes by thresholding the *p*-value (the default threshold is 1%).

2.3.5 Use of Peakfinding

A software package called Mpeak has been developed to implement the model-based peakfinding strategy described in the previous sections. The software and source code can be downloaded from http://www.chiponchip.org.

Given the assumptions of the peakfinding approach, Mpeak is ideally used for ChIP-chip experiments involving sequence-specific DNA binding factors or factors that generally localize at punctate binding sites such as TAF1. .

2.4 High-Level Analysis

We provide a cursory discussion of high-level analysis for ChIP-chip data because analysis is typically tailored to the biological question of interest motivating the ChIP-chip experiment. Standard questions for ChIP-chip with genomic tiling arrays include: (1) Genomic distribution of bound sites relative to known genes and transcripts, (2) Genomic position clustering of bound sites into dense domains, (3) Enrichment of any associated genes into relevant functional categories, (4), Sequence-conservation of bound

sites, (5) Sequence motifs characterizing bound sites, (6) Punctate or large spreads of binding enrichment, (7) Comparison with and co-localization of other factors by ChIP-chip, and (8) Correlation with matching transcription profiling information. In the following sections, we highlight some of the resources and strategies for addressing initial questions of annotation and visualizing genomic distribution. Chapters 3 and 4, in particular the Methods sections, will present more detailed strategies for high-level analysis.

2.4.1 Annotation

The central repositories for genomic annotation are the National Center for Biotechnology Information (NCBI) database and the University of California Santa Cruz (UCSC) Genome Browser[93,94]. Both resources contain sequence information for available genomes and assemblies. In addition, both sites contain coordinates and information for known genes, mRNA, and expressed sequence tags (ESTs), as well as annotation for known sites of genetic variation such as single-nucleotide polymorphisms (SNPs). The UCSC Genome Browser also has extensive annotation with respect to cross-species conservation, such as alignments of any given genomic region with other available species as well as estimates of the conservation rate over that region.

There is a variety of publicly available software and web resources for motif-analysis and gene set annotation developed concomitant with the widespread use of microarray gene expression data. These can be similarly applied for evaluating ChIP-chip binding sites [95-97]. An integrated web resource, *Cis*-Regulatory Element Annotation System (CEAS), is especially designed for ChIP-chip data annotation, given a set of

coordinates for ChIP-enriched regions it provides summaries of genomic distribution relative to known genes, conservation scores, and over-represented known transcription factor binding sites from TRANSFAC[98]. Although we do not use this tool for the work to be described, we believe that it can be useful for preliminary ChIP-chip data analysis.

2.4.2 Visualization

Graphical browsing of ChIP-chip data is enabled by software such as SignalMap from NimbleGen and the Integrated Genome Browser (IGB) from Affymetrix. ChIP-chip data files for browsing contain chromosomal coordinates and the ChIP log ratio for each probe. Visualization allows for initial identification of different patterns of ChIP enrichment, from punctate sites to large spans. Genomic distribution of ChIP enrichment relative to known genes or other functional elements can be initially assessed by simultaneous viewing of annotation files on additional tracks. Uploading custom tracks to the UCSC Genome Browser in designated BED, GFF, or WIG formats can also be used to visualize selected ChIP enriched regions in the context of all available annotation for transcripts, CpG islands, conservation, genetic variation, etc. (Figure 2-8) Details for the file formats are given on the UCSC Genome Browser[93].

Visualization and pattern classification tools developed for expression analysis such as TreeView and Cluster 3.0 are also useful for examining patterns of ChIP enrichment over a set of parallel loci, especially when evaluating tiling array data. For instance, visualization of TAF1, Pol II, H3Ac and H3K4me2 enrichment over windows centered on TAF1 sites reveals the co-localization of these marks over promoter regions [99-101] (Figure 2-9).

2.5 Data Management

Raw and pre-processed ChIP-chip data as well as resulting target lists and annotation are ideally stored within a database management system (DBMS). We use the open-source MySQL DBMS to facilitate data storage, retrieval, and manipulation. Indexing, by sub megabase intervals of genomic positions for all ChIP-chip data and genomic annotation tables, is critical to speed up comparisons and retrieval of specified genomic windows.

Submission to public repositories for microarray data such as ArrayExpress and Gene Expression Omnibus (GEO) are required for ChIP-chip publications. Data submission formats such as MAGE-ML, MIAME, and SOFT require detailed annotation of the ChIP-chip experiments [102,103].

Acknowledgments

Chapter 2 in part quotes sections from the following publications: (1) Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press. (2) Zheng, Ming; Barrera, Leah; Ren, Bing; Wu, Yingnian. ChIP-chip: data, model, and analysis. Biometrics, in press. I was a a secondary author and researcher in these works. For (1), I wrote the sections dealing with ChIP-chip data analysis. For (2), I contributed to the development of the model, testing of the algorithm, and in editing the manuscript.

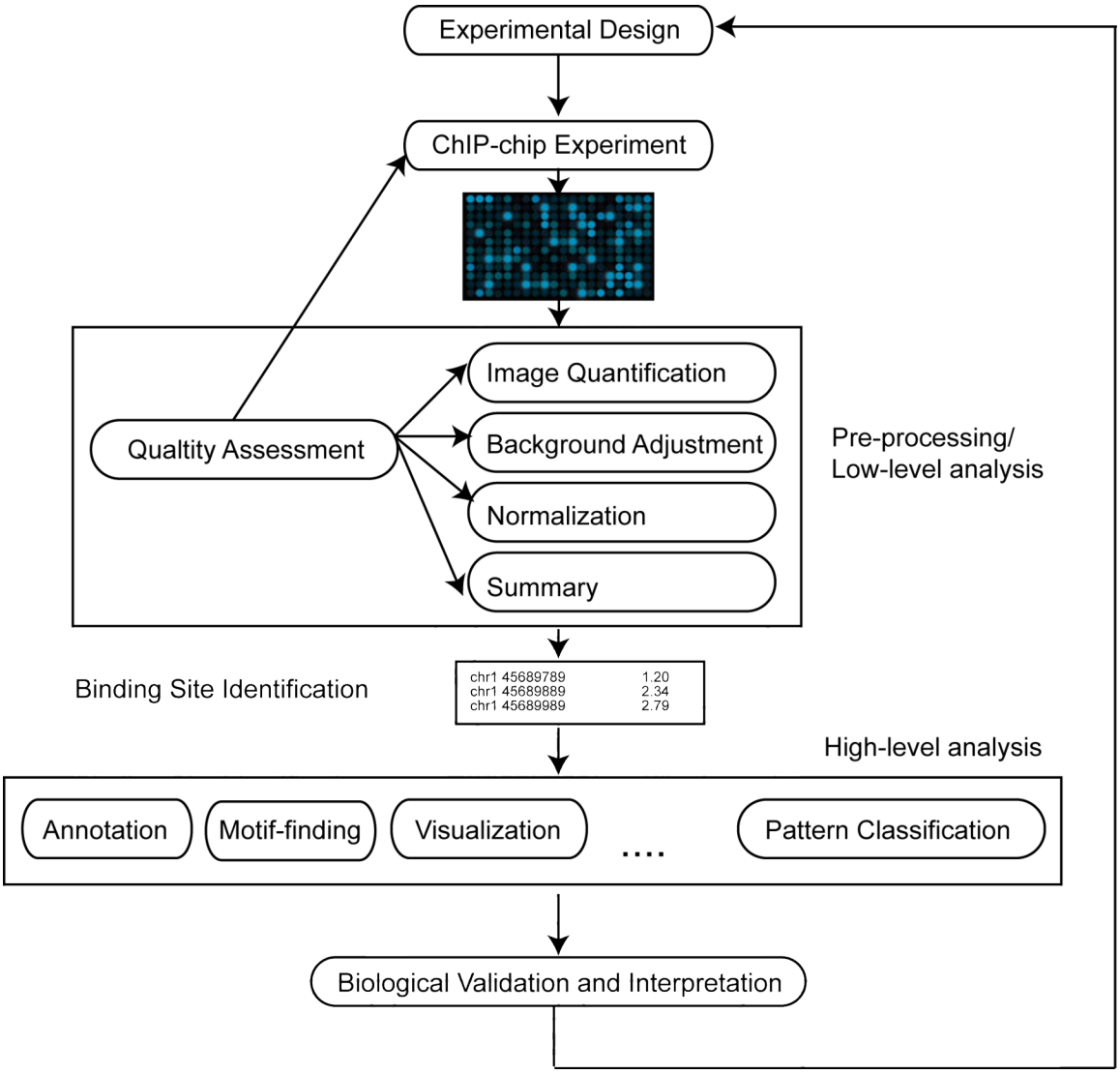


Figure 2-1. ChIP-chip analysis workflow.

(Figure inspired by expression analysis workflow presented at the Cold Spring Harbor Systems Biology Workshop by Dr. Xiaoyue Zhao).

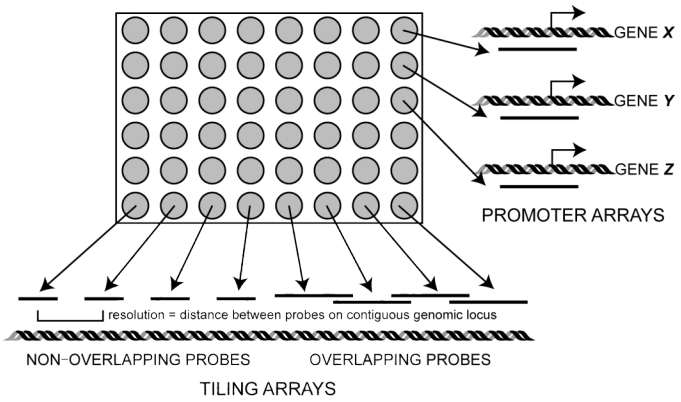
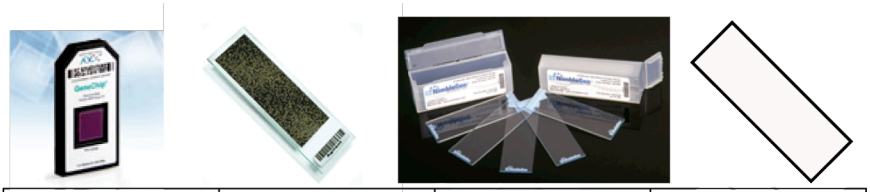


Figure 2-2. Coverage and resolution of arrays.

(From publication: Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press.)



	Affymetrix	Agilent	NimbleGen	PCR array (example)
Probe length (bp)	25	60	50	~600
Tiling resolution (bp)	35	100	100	1,000
Coverage (Human genome)	Genome, Promoters	Custom, Genome, Promoters	Custom, Genome, Promoters	Custom, Promoters
Features/Array	6,500K	244K	390K	25K
Software Support	GeneChip Operating Software (GCOS) Tiling Analysis Software (TAS) Integrated Genome Browser (IGB)	ChIP Analytics	SignalMap	In-House
Academic Software	Model-based analysis of tiling arrays (MAT), TileHGMM, TileMAP	Model-based analysis of 2-color arrays (MA2C), Mpeak, ChIP On Tiled Arrays (ChIPOTle), Joint Binding Deconvolution (JBD)	Mpeak, MA2C, ChIPOTle	Single-Array Error Model (SAEM)
Web Resources	UCSC Genome Browser *Cis*-regulatory Element Annotation System (CEAS) Promoter Array Analysis Server (for NimbleGen Promoter Arrays) TAMALPAIS (for NimbleGen Tiling Arrays) TileScope			

3/23/2007

Figure 2-3. ChIP-chip microarray platforms.

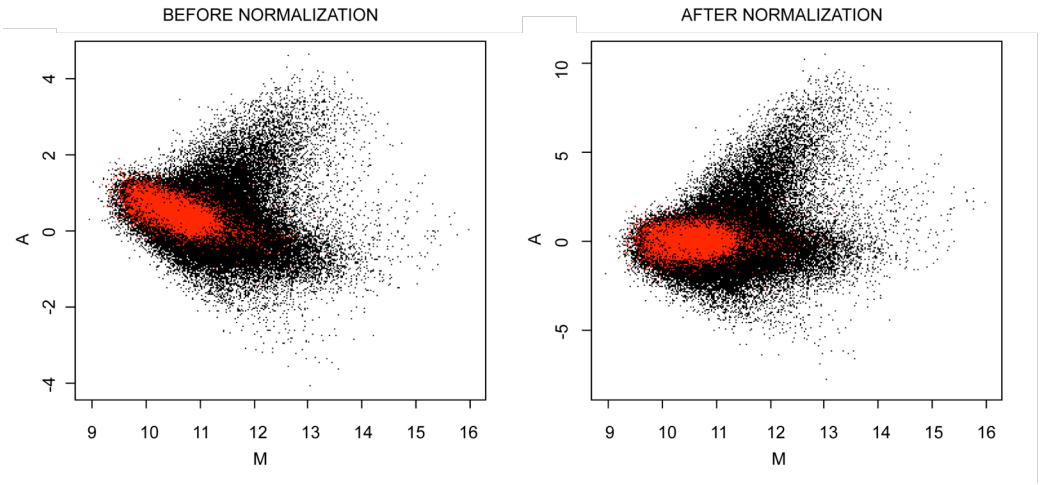


Figure 2-4. M vs. A plot before and after normalization.

Control probes used to calculate loess normalization fit are highlighted in red. (From publication: Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press.)

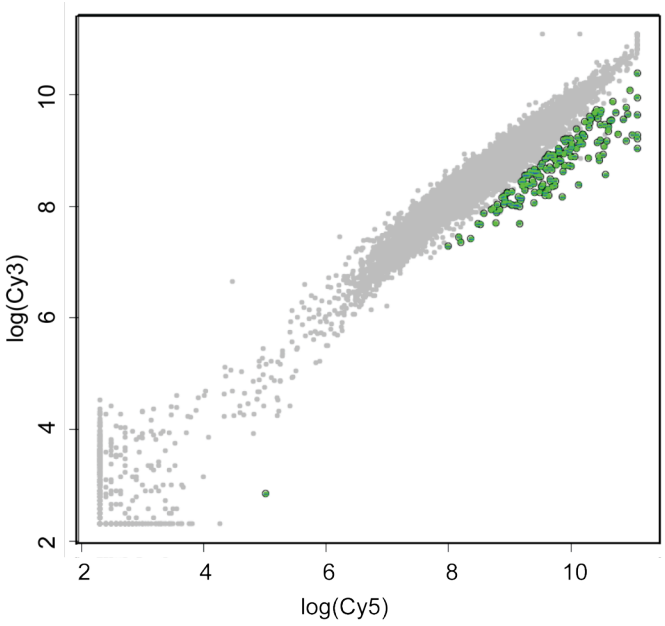


Figure 2-5. Scatterplot of Cy5 versus Cy3 enrichment.

Green spots indicate enriched probes based on the Single-Array Error Model. (From publication: Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press.)

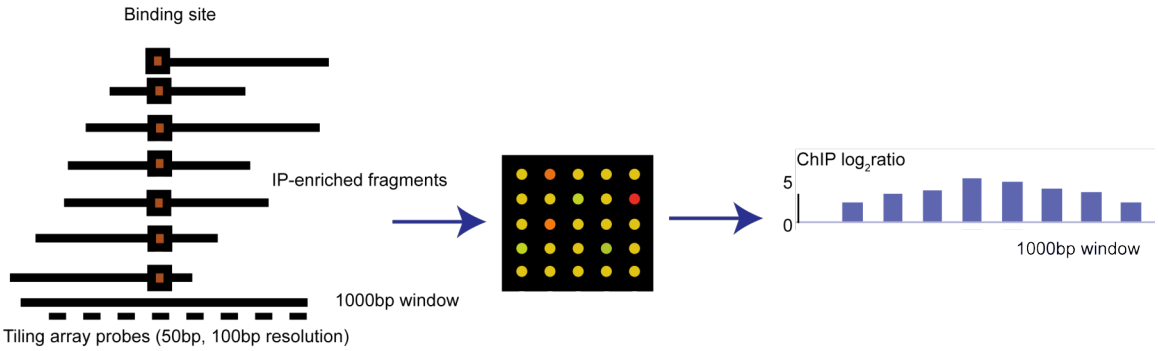


Figure 2-6. ChIP-chip enrichment with high-resolution tiling arrays.

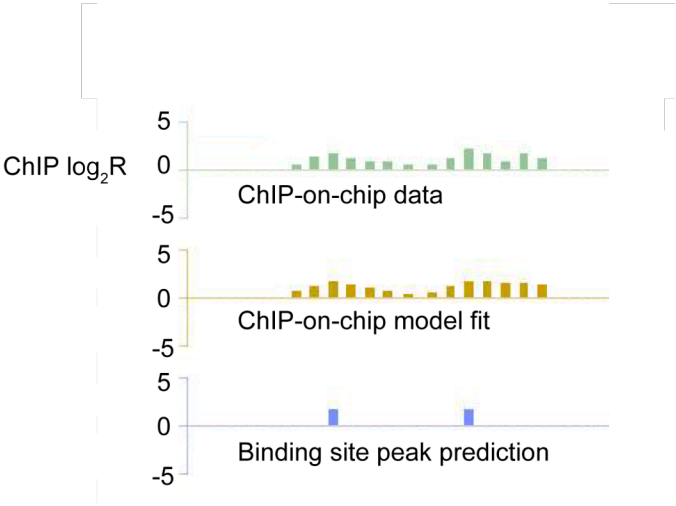


Figure 2-7. Example of binding site resolution by peakfinding.

(From publication: Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press.)

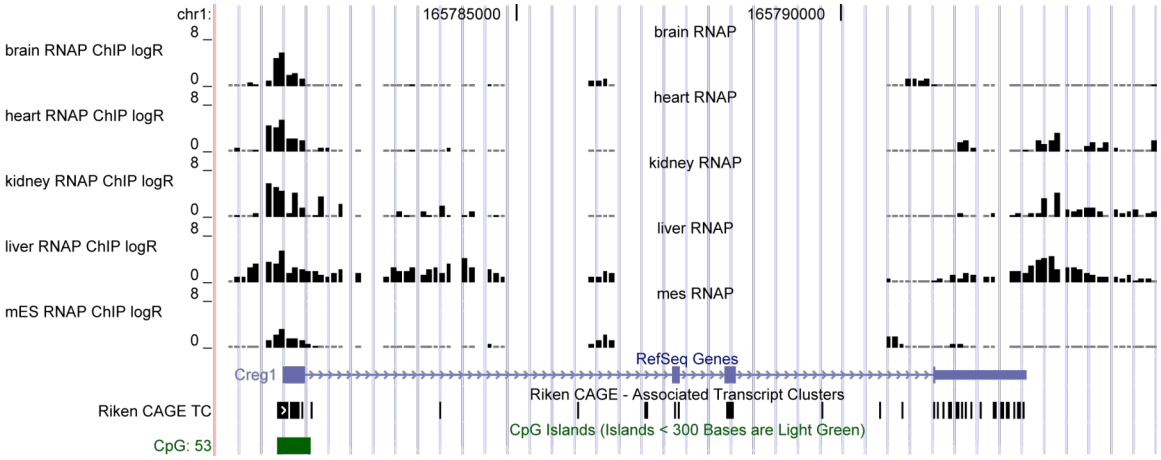


Figure 2-8. UCSC Genome Browser screen shot.

(From publication: Kim, Tae H; Barrera, Leah; Ren, Bing. Genome-wide analysis of protein binding in mammalian cells. Current Protocols in Molecular Biology, in press.)

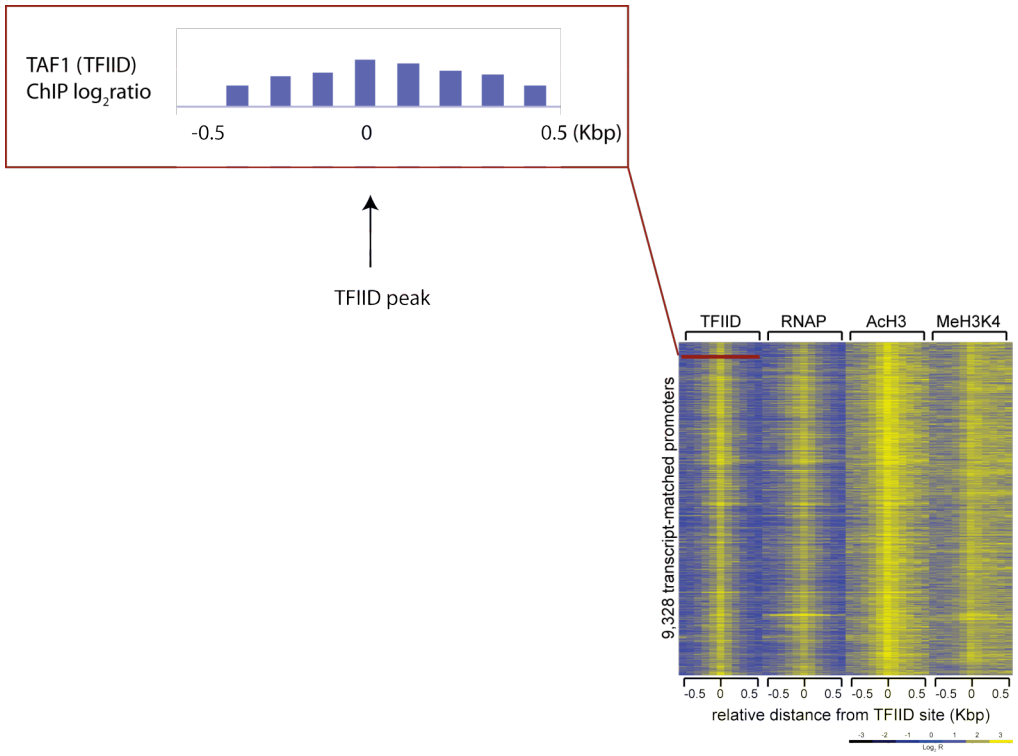


Figure 2-9. ChIP-chip profiles with TreeView.

(Adapted from publication: Kim, Tae H; Barrera, Leah O; Zheng, Ming; Qu, Chunxu; Singer, Michael A.; Richmond, Todd A.; Wu, Yingnian; Green, Roland; Ren, Bing. A high-resolution map of active promoters in the human genome. Nature, Vol.436, 2005).

Chapter 3

A High-resolution Map of Active Promoters in the Human Genome

In eukaryotic cells, transcription of every protein-coding gene begins with the assembly of an RNA Polymerase II (Pol II) preinitiation complex (PIC) on the promoter[12]. The promoters, in conjunction with enhancers, silencers and insulators, define the combinatorial codes that specify gene expression patterns[104]. Our ability to analyze the control logic encoded in the human genome is currently limited by a lack of accurate information of the promoters for most genes[105]. Here, we describe a genome-wide map of active promoters in human fibroblast cells, determined by experimentally locating the sites of PIC binding throughout the human genome. This map defines 10,567 active promoters corresponding to 6,763 known genes and at least 1,196 un-annotated transcriptional units. Features of the map suggest extensive usage of multiple promoters by human genes and widespread clustering of active promoters in the genome. In addition, examination of the genome-wide expression profile reveals four general classes of promoters that define the transcriptome of the cell. These results provide a global view of the functional relationship among the transcriptional machinery, chromatin structure and gene expression in human cells.

3.1 Promoter Mapping in Human Fibroblast Cells

3.1.1 Overview of Strategy

The PIC consists of the RNA Polymerase II (Pol II), the transcription factor IID (TFIID) and other general transcription factors[17]. Our strategy to map the PIC binding sites involves chromatin immunoprecipitation coupled DNA microarray analysis (ChIP-chip), which combines the immunoprecipitation of PIC-bound chromatin from formaldehyde crosslinked cells with parallel identification of the resulting bound DNA sequences using DNA microarrays[40,81]. Previously, we have demonstrated the feasibility of this strategy by successfully mapping active promoters in 1% of the human genome that correspond to the 44 genomic loci known as the ENCODE regions[81,106]. To apply this strategy to the entire human genome, we fabricated a series of DNA microarrays[82] containing roughly 14.5 million 50-mer oligonucleotides, designed to represent all the non-repeat DNA throughout the human genome at 100 basepairs (bp) resolution. We immunoprecipitated TFIID-bound DNA from the primary fibroblast IMR90 cells with a monoclonal antibody that specifically recognizes the TAF1 subunit of this complex (TBP associated factor 1, formerly $TAF_{II}250$[107], Figure 3-1). We then amplified and fluorescently labeled the resulting DNA, and hybridized it to the above microarrays along with a differentially labeled control DNA (Figure 3-1 A).

3.1.2 Summary of TFIID Binding and Annotation

We determined 9,966 potential TFIID-binding regions using a simple algorithm requiring a stretch of four neighboring probes to have a hybridization signal significantly

above the background. To independently verify these TFIID-binding sequences, we designed a condensed array that contained a total of 379,521 oligonucleotides to represent these sequences and 29 control genomic loci selected from the 44 ENCODE regions[106] at 100 bp resolution. ChIP-chip analysis of two independent samples of IMR90 cells confirmed the binding of TFIID to a total of 8,597 regions, ranging in size from 400 bp to 9.8 Kbp (Figure 3-1 B). We further defined a total of 12,150 TFIID-binding sites within the 8,597 fragments using a peak finding algorithm that predicts the most likely TFIID-binding sites based on the hybridization intensity of consecutive probes with significant signals (Figure 3-1 A).

Next, we matched these 12,150 TFIID-binding sites to the 5' end of known transcripts in three public transcript databases (DBTSS[108], RefSeq[109], GenBank human mRNA collection[110]) and the EnsEMBL gene catalog[111]. To account for the uncertainty of our knowledge of the true 5' end of transcripts and the uncertainty of predicted TFIID-binding positions due to noise within the microarray data, we chose an arbitrary distance of 2.5 Kbp as a measure of close proximity. We found that 10,553 (87%) TFIID-binding sites were within 2.5Kbp of annotated 5' ends of known mRNA. We resolved common TFIID-binding sites mapping to similar 5' ends to define a non-redundant set of 9,328 5'end-matched TFIID-binding sites. Of these TFIID-binding sequences 7,789 (83%) were found within 500 bp of the putative transcription start sites (TSS) (Figure 3-1 C). Since these 9,328 DNA sequences were bound by TFIID *in vivo* and within close proximity to the 5'end of known transcripts, we defined them as promoters for the corresponding transcripts (Data Table 3-1).

Of these 9,330 promoters, 8,960 were mapped within 2.5 Kbp of the 5' end or within annotated boundaries of 6,763 known genes in the EnsEMBL gene catalog[111] (Figure 3-1 D, Data Table 3-1). The remaining 368 promoters corresponded to transcripts not contained within these boundaries of EnsEMBL genes, and therefore provide support for inclusion of these transcripts to the current gene catalogs. The list of promoters also confirmed 5,118 previously annotated promoters[108], and defined 4,210 new promoters for at least 2,627 genes (Figure 3-1 E, Data Table 3-1).

3.1.3 Independent Support and Characterization of TFIID Sites

Four independent analyses validated the high specificity and accuracy of the active promoters detected in IMR90 cells. First, ChIP-chip analysis using an anti-Pol II antibody (8WG16) confirmed the binding of Pol II to at least 9,050 (97%) of the 9,328 promoters in IMR90 cells (Figure 3-2). Second, standard chromatin immunoprecipitation (ChIP) performed on 28 promoters randomly selected from the above list confirmed the occupancy of Pol II on all but one promoter (Figure 3-3). Third, the 9,328 active promoters are enriched for known promoter-associated sequences such as CpG islands, and the INR and DPE core promoter elements (Figure 3-1 F). The percentage of CpG-associated promoters (88%) was significantly higher than the previous estimate (56%)[112], suggesting that CpG islands may play a more general role in gene expression than previously appreciated. Surprisingly, we did not find the TATA box to be significantly enriched in these promoters (Figure 3-1 F). This may be due to a lack of conservation of the TATA box in human promoters; alternatively, this may indicate that the TATA box is not a general promoter motif for human genes. This observation is in line with previous

reports that the TATA box is only present in a small number of promoters in yeast and in Drosophila[113]. Fourth, ChIP-chip analysis using antibodies that recognize acetylated histone H3 (H3ac) or di-methylated lysine 4 on histone H3 (H3K4me2) showed that over 97% of the 9,328 promoters were associated with these known epigenetic marks for active genes (Figure 3-4)49. Interestingly, the localization of H3K4me2 in these promoters was predominantly downstream of the TFIID-binding site (Figure 3-4), and the mechanisms for such chromatin organization at human promoters are currently unclear.

3.1.4 Novel Promoters

Among the 12,150 mapped TFIID-binding sites, 1,597 are found more than 2.5 Kbp away from previously defined 5'ends of mRNA, and may represent promoters for novel transcripts or genes (Data Table 3-2). Of these, 607 non-redundant TFIID-binding sites were matched within 2.5 Kbp of the 5' ends of the Expressed Sequence Tag (EST)-based gene models, indicating that they may indeed produce mRNA (Data Table 3-2). The remaining TFIID-binding sites were further filtered to a set of 632 putative promoters by requiring the occupancy of Pol II and presence of H3ac and H3K4me2 within 1 Kbp of these sites (Figure 3-5).

To verify that these promoters drive transcription, we analyzed mRNA from the IMR90 cells, using 50-mer oligonucleotide arrays that represent a 28 Kbp sequence surrounding 567 of 632 unmatched putative promoters. At least 35 novel transcription units were identified near the putative promoter regions, suggesting that these may represent new transcription units yet to be annotated in the human genome (Data Table 3-3). The failure to detect mRNA from the other putative promoters may indicate that

these transcripts are highly unstable. Indeed, at least one putative promoter is located within 250 bp upstream from a predicted miRNA[114] (Figure 3-6), suggesting that some putative promoters could transcribe non-coding RNA that might have escaped detection by conventional mRNA isolation techniques.

In all, we defined a set of 1,239 putative promoters that correspond to previously un-annotated transcription units (Figure 3-5,Data Table 3-2). Evolutionarily conserved regions were found in a majority of these putative promoters (Figure 3-7). In addition, they were significantly enriched for core promoter motifs including INR (46%) and DPE (40%) and overlapped with CpG islands (40%,Figure 3-8). These results suggest that many of the putative promoter sequences that we have defined by TFIID-binding sites may indeed be functional promoters. There are 828 putative promoters located in the intergenic regions. These promoters, together with the 368 promoters that matched to transcripts outside the EnsEMBL genes, may suggest the existence of 1,196 novel transcription units outside the current gene annotation[115]. This number corresponds to about 13% of the 8,960 promoters that were matched to known genes; therefore, we estimate that there are likely additional 13% of the human genes that remain to be annotated in the genome. This number agrees well with a recent estimate of the total number of human genes[115], but is considerably lower than estimates based on number of transcripts detected by microarrays, SAGE, and other methods[92,116-118]. It is conceivable that promoters for many low-abundance transcripts may be infrequently occupied by TFIID and possibly escaped detection by our assays. Alternatively, it is possible that the novel transcripts detected by the other studies are products from a different transcription machinery or process.

3.2 Features of Active Promoters

3.2.1 Clustering of Active Promoters

Two notable features were apparent in this map of active promoters. First, large domains of four or more consecutive genes were found to be simultaneously bound by PIC and likely transcribed in the IMR90 cells. At least 256 clusters, consisting of 1,668 EnsEMBL genes, can be classified into such regions, and the number of clustered promoters is highly significant ($p << 0.001$, Data Table 3-5). The clustering of active promoters is consistent with previous findings that co-regulated genes tend to be organized into coordinately regulated domains[119-122].

3.2.2 Alternative Promoter Usage

Second, a large number of genes contained two or more active promoters (Data Table 3-4). In general, these multiple promoters correspond to transcripts with either different 5' UTR sequences or distinct first exons (i.e., *PTEN*) but do not affect the open reading frames. In some cases, however, distinct proteins were produced from multiple promoters (i.e., *NR2F2*, *WEE1*). In other cases**,** transcripts undergo differential splicing and polyadenylation (i.e., *NFKB2*, *STAT3*). The widespread usage of multiple promoters in this single cell type indicates a greater complexity of the cellular proteome than previously expected and also reveals highly coordinated regulation of transcriptional initiation, splicing, and polyadenylation throughout the genome[123]. To experimentally verify our observations regarding multiple promoter utilization in IMR90 cells, we selected the *WEE1* gene for further analysis. Two TFIID-binding sites were mapped

within this gene, corresponding to the 5' ends of two distinct mRNAs, NM_003390 and AK122837 (Figure 3-9 A). Each mRNA encodes a distinct protein: one encodes a well-characterized full length version of WEE1 protein, and the other only the kinase domain. We detected both transcripts in a steady state, asynchronous population of IMR90 cells (Figure 3-9 B). Interestingly, the shorter transcript appears to be most abundant in G0 phase, while the longer transcript is highly transcribed in both G0 and S phase (Figure 3-9 C), suggesting that the two promoters in the *WEE1* gene may have distinct cell cycle functions.

3.3 Comparison with Expression Profiling

The active promoter map in IMR90 cells allowed us to systematically investigate the functional relationship between the transcription machinery and gene expression. We examined the genome-wide expression profiles of IMR90 cells and correlated the expression status of 14,437 EnsEMBL genes to promoter occupancy by the PIC.

3.3.1 PIC Binding and Expression

The comparison revealed four general classes of genes (Figure 3-10, Data Table 3-6). Class I consists of 4,415 genes whose promoters were bound by the PIC, and transcripts were detected. Class II includes 658 genes whose promoters were bound by the PIC, but no transcript was detected. Class III contains 2,879 genes that were transcribed in IMR90 cells but the PIC was not detected on their promoters. Class IV comprises of the remaining 6,485 genes whose promoters were not bound by PIC and their corresponding transcripts were not detected.

The genes in class I and class IV, representing over 75% of the genes examined, support the general model that formation of the PIC on the promoters leads to transcription. The class II and III genes, on the other hand, are inconsistent with this model and may indicate other mechanisms responsible for the expression of these genes. We postulate that the discrepancy between the PIC formation and transcription on the class II promoters are due to at least two possibilities. The first possibility is that the PIC assembles on these promoters, but the PIC formation is not sufficient to initiate transcription. Additional regulatory steps, such as promoter clearance or elongation may be rate-limiting in transcription of these genes[124]. Some notable examples in class II are the immediate early genes, *FOS* and *FOSB*; the heat shock protein genes, *HSPA6* and *HSPD1*; and the DNA damage repair genes, *MSH5* and *ERCC4*. The second possibility is that transcription actually takes place at these promoters, but the resulting mRNAs are post-transcriptionally degraded, as in miRNA-mediated post-transcriptional silencing[125].

In contrast to class II, genes in class III appear to be transcribed, but the PIC binding on their promoters was not detected. This could simply be due to moderate sensitivity of our method81. To address this issue, we performed standard ChIP assay to detect binding of TFIID and Pol II on 10 randomly selected class III gene promoters. Nearly 60% of the promoters were weakly associated with TFIID and Pol II in these cells, and were marked by enrichment ratios less than 2-fold but nonetheless above the observed background (Figure 3-3). Hence, the failure to detect TFIID and Pol II occupancy in roughly 60% of the class III promoters (~1,700) may be due to weak signals that fall below the detection sensitivity of our method. This result indicates that the promoters of a significant fraction of class III genes are open and accessible for

transcription, but PIC assembles on these promoters transiently, weakly or only during the early stage of fibroblast differentiation.

3.3.2 Histone Modifications and Expression

In order to understand the functional relationship between the histone modification status and gene expression, we examined the histone modifications (H3ac and H3K4me2) in 29 ENCODE regions[106] (Data Table 3-7), with a specific focus on the four classes of gene promoters. As expected, these epigenetic markers were associated with virtually all class I and class II genes, and the vast majority of class III genes. However, roughly 20% of the class IV genes were also associated with these markers (Figure 3-10). This result suggests that a significant number of genes not actively transcribed are also associated with these epigenetic markers. We speculate that these histone modifications may serve to restrict genome expression potential and define the transcriptome capacity of the cell, and the transcription regulators and machinery collaborate with these epigenetic markers to further restrict the transcriptome to generate a unique pattern of genome expression.

3.4 Conclusion

Our results provide an initial framework for analysis of the cis-regulatory logic[126] in human cells. The high-resolution map of active promoters in IMR90 cells will enable detailed analysis of transcription factor binding sites within these regions. The promoter map described here can also serve as a reference to understand gene expression in other cell types. We expect that a survey of additional cell types using the same approach will

allow comprehensive mapping of all promoters in the human genome, and help elucidate the control logic that governs gene expression in different cell types in the body.

3.5 Methods

Experimental design

To identify active promoters in human cells, we isolated DNA bound by the general transcription factor IID (TFIID) from crosslinked primary fibroblast IMR90 cells by chromatin immunoprecipitation with an antibody that specifically recognizes the TAF1 subunit of this complex (TBP associated factor 1, formerly TAF250). The enriched DNA was then amplified and fluorescently labeled, and hybridized to the high-density oligonucleotide arrays along with a differentially labeled control DNA. We determined the potential TFIID binding sites using a simple statistical threshold requiring a stretch of four neighboring oligos to have a hybridization signal significantly above background. A total of 9,966 clusters of TAF1 binding sites were identified by this analysis. To verify the binding of TAF1 to these sequences, we designed a new array that contains a total of 379,521 50-mer oligos to represent the 9,966 putative TAF1 binding sequences plus 29 control genomic loci (selected from the ENCODE regions, each ranging in size from 500 Kbp to 1.9 Mbp) at 100 bp resolution. TAF1 bound DNA was isolated from two independent samples of IMR90 cells and was labeled and hybridized to these arrays. A total of 8,597 TAF1 binding regions, ranging in size from 400 bp to 9.8 Kbp, were confirmed by the replicate experiments.

Samples used, extract preparation and labeling

IMR90 cells were grown and maintained according to the direction from American Type Culture Collection. Cells were harvested and crosslinked with 1% formaldehyde when they reached ~80% confluency on the plates. Chromatin immunoprecipitation was performed as described previously (http://www.pnas.org/cgi/data/1332764100/DC1/1), with the following modifications[64]. Antibodies for ChIP were obtained from commercially available sources: mouse monoclonal antibody against Pol II (catalog # MMS-126R, Covance), mouse monoclonal TAF1 antibody (catalog # sc-735, Santa Cruz Biotechnology), and rabbit polyclonal H3ac and H3K4me2 (catalog # 06-599 and 07-030 respectively, Upstate). Following the ligation mediated PCR (LM-PCR) step, additional PCR reactions were performed to generate 100 μg of ChIP DNA for hybridization. 200 ng of LM-PCR products were amplified for five additional cycles under the same LM-PCR conditions.

One microgram (μg) of LM-PCR products were used for labeling and hybridization to each array. One microgram of immunoprecipitated or total genomic LM-PCR DNA was mixed with 40 μL of 1 μM Cy5 or Cy3 end labeled random prime nonamer oligonucleotides (TriLink Biotechnologies) respectively with the bacterial label control DNA in a total volume of 88 μL. The DNA and random primers were annealed by heating the sample to 98°C for 5 minutes and chilled quickly in ice water for 2-3 minutes. Two microliter of (100 units) of E. coli DNA polymerase Klenow fragment and 10 μL of 10 mM equimolar mixture of dATP, dTTP, dCTP, and dGTP were added to the annealed DNA sample and incubated at 37°C for 2 hours. The reaction was stopped by addition of 10 μL of 0.5 M EDTA. The labeled sample was ethanol precipitated by addition of 11μL 5 M NaCl and 110 μL isopropanol. The precipitate was collected by

centrifugation and the resulting labeled DNA pellet was washed with 80% ethanol (V/V). The pellet was dried under vacuum for 5-15 minutes to remove any remaining liquid, and the resulting dry labeled DNA pellet was resuspended in 10 μL dH_2O.

Hybridization procedure and parameters

Equal amounts (12 μg) of Cy5 and Cy3 labeled DNA samples were mixed, and 4 μL 2.94 nM Xenohybe control oligos (an equimolar mixture of 5'TTGCCGATGCTAACGACGCATCAGACTGCGTACGCCTAAGCAACGCTA3' and 5'CATTGCTGTGCGTACGCAGTCAAGTCGATCACGCTAACTCGTTGCGAC3') was added to the mixture. The sample was vacuum dried under low heat until the volume of sample was less than 14.4 μL. The final volume of DNA was adjusted to 14.4 μL with dH_2O. To this sample, 11.25 μL 20X SSC, 18 μL 100% formamide, 0.45 μL 10% SDS, 0.45 μL 10X TE (100mM Tris, 10mM EDTA), and 0.45 μL equimolar mixture of Cy3 and Cy5 labeled CPK6 oligonucleotides (5'TTCCTCTCGCTGTAATGACCTCTATGAATAATCCTATCAAACAACTCA3' and 5'TTCCTCTCGCTGTAATGACCTCTATGAATAATCCTATCAAACAACTCA3', respectively) were added to prepare the hybridization mixture. The hybridization sample was heated to 95 °C and was applied to the slide and incubated in the MAUI® Hybridization Station (BioMicro Systems, Inc.) at 42 °C for 16-20 hours.

The hybridized slides from the MAUI® Hybridization Station were washed once in Wash 1 (0.2X SSC, 0.2% SDS, 0.1 mM DTT) for 10-15 seconds and followed by another wash in Wash 1 (0.2X SSC, 0.2% SDS, 0.1 mM DTT) for 2 minutes with gentle agitation. The slides were then washed in Wash 2 (0.2X SSC and 0.1mM DTT) for 1

minute and followed by a wash in Wash 3 (0.05X SSC and 0.1 mM DTT) for 15 seconds. The slides were dried by centrifugation.

Measurement data and specifications

The hybridized arrays were scanned on an Axon GenePix 4000B scanner (Axon Instruments Inc.) at wavelengths of 532nm for control (Cy3), and 635nm (Cy5) for experimental sample. Data were extracted from the scanned images using the NimbleScan 2.0 program (NimbleGen Systems, Inc.). The arrays were gridded using the automated gridding algorithm, and extracted in two channels using a mean intensity calculation of the interior of the gridded rectangular features upon extraction, and each pair of N probe signals were converted into a scaled log ratio using the function:

R(i) = Log (Experimental(i) / Control(i))

The raw microarray data can be visualized by SignalMap software (Nimblegen Inc.).

Array design

The 38 genome scan arrays contained a total of 14,535,659 50-mer oligonucleotides, positioned at every 100 basepairs (bp) throughout the human genome as described in the oligonucleotide array description files.

A condensed array used to verify the results from genome scan arrays contained a total of 379,521 oligonucleotides to represent the 9,966 putative TAF1 binding sequences plus 29 control genomic loci (selected from the ENCODE regions, each ranging in size from 500 Kbp to 1.9 Mbp) at 100 bp resolution.

Initial identification of TAF1 binding regions

After scanning and image extraction, Cy5 (TAF1 IP) and Cy3 (input) signal values for each of the 38 arrays tiling the non-repetitive sequences of the human genome

at 100 bp resolution (NCBIv34) were normalized by intensity-dependent Loess[127]. Median filtering (window size=3 probes) was used to smooth logR (Cy5/Cy3) data across the tiled regions. For each array, IP-enriched probe clusters were defined as regions with a minimum of 4 probes separated by a maximum of 500 bp with filtered logR greater than 2.5 standard deviations from the mean log ratio.

Peak finding and confirmation of TAF1 binding sites

To confirm evidence of TAF1 binding at the initial 9,966 IP-enriched clusters, dye-swapped replicate ChIP experiments were hybridized to a custom array tiling only these putative binding regions (each extended upstream and downstream by 1.5 times its length) and the 29 ENCODE[19] regions as control, at 100 bp resolution. Measured intensities for dye-swapped replicates were median-scale normalized and corresponding log ratios averaged. The log ratios across the regions were visualized using SignalMap. Given the 100 bp resolution of the arrays, we developed a peakfinding model to more precisely define binding sites (Chapter 2). In this original implementation of the peakfinding strategy, the significance of a specific peak $\hat{P}$ was based on the *p*-value for the following hypothesis test:

H_0 : the signals around $\hat{P}$ are generated by Gaussian noise
H_1 : the signals around $\hat{P}$ are not generated by Gaussian noise

Let μ_{noise} and σ^2_{noise} represent the mean and variance of all the signals covered by the fitted triangle centered at $\hat{P}$, n the number of signals, and $\hat{\sigma}^2$ variance of the residuals to the model fit. Then the likelihood-ratio test statistic is:

$$-2\log\frac{likelihood(\text{Gaussian noise assumption})}{likelihood(\text{model fitting})} = n\cdot(\log(\sigma^2_{noise}) - \log(\hat{\sigma}^2)) \sim \chi^2_3$$

The degrees of freedom equal 3 because we fit three additional parameters in our model (one α and two β's because of the asymmetry of the triangles). We used a significance threshold $P \leq 0.2$ to define peaks as binding sites.

Classifying promoters by matching to 5' end and gene annotation

We compared the location of 12,150 TAF1 binding sites to annotated 5'end of transcripts from RefSeq, GenBank, and DBTSS. RefSeq transcript (refGene.txt) and GenBank mRNA (all_mrna.txt) coordinates were downloaded from UCSC Genome Browser (http://genome.cse.ucsc.edu) in Sept. 2004 (HG16, July2003/NCBI Build 34). The set of GenBank human mRNA data (all_mrna.txt) was filtered to include only mRNA alignments to the genome which match 95% of the transcript length. DBTSS data was downloaded from DBTSS Home (http://dbtss.hgc.jp) in Jan. 2004. HG13 DBTSS coordinates were converted by blat alignment of promoter sequences to HG16 assembly from the UCSC Genome Browser.

5' End Annotation Source	**Number of Transcripts**
DBTSS	8,793
RefSeq	22,074
GenBank mRNA	118,346

We found 10,504 binding sites within 2.5 Kbp of an annotated 5'end from DBTSS, RefSeq, or GenBank. To remove possibly redundant matches, we matched binding sites to their closest 5' ends. For each 5' end matched, the closest TAF1 binding site was selected to define a non-redundant set of **9,281** transcript-matched promoters.

We then used MegaBLAST[128] to match these 9,281 transcript-matched promoters to 22, 222 EnsEMBL[111] annotated genes (EnsEMBL v26) by requiring that the matched transcript have a 95% sequence identity and a hit length of at least 50 bp with the best aligning EnsEMBL transcript. Additionally we verified that the matching transcript is within the annotated boundaries of the matching EnsEMBL gene. By this strategy we matched 7,920 transcript-matched promoters to 6,197 genes with high-confidence. 216 additional transcript matches to 206 EnsEMBL genes were made using translation tables knownToEnsembl and ensGeneXref downloaded from UCSC Genome Browser in Feb. 2005 (HG17, May2004 /NCBI Build 35). To define the overlap of EnsEMBL genes with our transcript-matched promoters, the 1,145 transcript-matched promoters not matched to EnsEMBL genes by strategies described above were matched to corresponding EnsEMBL genes if they fall within the annotated gene boundaries.

The 1,646 TAF1 binding sites (12,150 minus 10,504) outside 2.5 Kbp of the annotated 5' ends from DBTSS, RefSeq, and GenBank were then matched to the 5' ends of Acembly gene models based on EST clustering. Acembly coordinate information (acembly.txt) for 222,699 genomic alignments was downloaded from the UCSC Genome Browser in Sept. 2004 (HG16, July2003/NCBI Build 34). 749 promoters mapped within 2.5 Kb of Acembly annotated 5' ends were filtered to a set of **646** non-redundant set of Acembly-matched promoters. **39** of these Acembly-matched promoters were found to be within 2.5Kbp of EnsEMBL annotated gene starts and added back to the list of transcript-matched promoters, resulting in final total of **607** Acembly-matched promoters.

The remaining 897 sites not matched within 2.5Kbp of annotated 5' ends from DBTSS, RefSeq, GenBank, and Acembly were filtered to define a set of **644** putative

promoters by requiring the identification of binding sites for Pol II and H3ac and H3K4me2 within 1 Kbp of the TAF1 sites. **10** of these promoters not matched by DBTSS, RefSeq, GenBank, and Acembly were found to be within 2.5Kbp of EnsEMBL annotated gene starts and added back to the list of transcript-matched promoters, resulting in a total of 634 filtered unmatched promoters. Two of the 634 correspond to chrY regions homologous to other hits and were thus removed from analysis to give a final count of **632** filtered unmatched promoters.

The combination of 9,281 transcript-matched promoters, 39 Acembly-matched promoters, and 10 putative promoters within 2.5Kbp of annotated ensEMBL genes represent a total of 9,330 promoters matched to known 5' ends. Two of 9,330 were found to correspond to chrY regions homologous to other hits and were thus removed from analysis to give a final count of **9,328** promoters matched to known 5' ends.

In the discussion, we partitioned the total of **10,567** promoters identified as **9,328** promoters matched to known 5' ends (transcript-matched promoters) and **1,239** putative promoters (**607** Acembly + **632** filtered unmatched).

A total of **8,960** of the **9,328** 5'end/transcript-matched promoters mapped to **6,763** EnsEMBL genes by alignment, location within the gene, or translation from annotation tables. To map transcript-matched promoters to nearby DBTSS annotated 5' ends, we checked whether these promoters were within 2.5 Kbp of the DBTSS annotated start site and that the matching transcript is on the same strand as the DBTSS-associated transcript. By such criteria, we found **5,118** promoters within 2.5 Kbp of **3,996** DBTSS annotated promoters.

The coordinates of the TAF1 binding sites for the set of putative promoters were also compared to the genomic coordinates of the EnsEMBL annotated genes to determine whether to classify their location as intergenic or within genes. Putative promoters were classified as intergenic if they fall outside the annotated start and end of all the genes in the EnsEMBL gene catalog. **411** of **1,239** putative promoters fall within EnsEMBL annotated genes while **828** fall outside of ensEMBL genes.

(Unless otherwise described, required conversions of coordinates between different assemblies/NCBI Builds were done using the utility *liftOver* and chain files from the UCSC Genome Browser).

Validation of promoters

Quantitative real-time PCR was performed with 0.5 ng of TAF1 or Pol II ChIP DNA and enriched total genomic DNA, as described previously[42,56]. The quantitative real-time PCR of each sample was performed in duplicate using iCycler™ and SYBR green iQ™ SYBR green supermix reagent (Bio-Rad Laboratories). The threshold cycle (Ct) values were calculated automatically by the iCycle iQ™ Real-Time Dectection System Software (Bio-Rad Laboratories). Normalized Ct (ΔCt) values for each sample were then calculated by subtracting the Ct value obtained for the unenriched DNA from the Ct value for the promoter DNA ($\Delta Ct = Ct_{promoter} - Ct_{total}$). The fold enrichment of the tested promoter sequence in ChIP DNA over the unenriched DNA was estimated as described previously[42,56]. Primers used for this analysis are list below.

Validation of putative promoters

An array containing 567 of the 632 putative promoters with no 5' end matches was designed by taking 13Kbp upstream and downstream sequences from each site and

tiling these regions at 38bp resolution with 49-mer probes (180,873 probes). Controls selected from ENCODE regions were also tiled in the array at the same resolution (200,378 probes). To this array, Cy5 labeled cDNA synthesized from total RNA extracted from IMR90 cells were hybridized. Hybridization was performed as described above.

To determine transcribed regions proximal to the putative promoters we used a strategy similar to that used to define transcriptionally active regions (TARs) with high-resolution tiling arrays [129]. Here, we required a minimum set of four consecutive probes exhibiting fluorescence intensities above the 90th percentile. Regions defined by these criteria as transcribed were then checked for overlap with known exons and/or location within genomic loci of known transcripts based on annotation from the knownGene table from the UCSC Genome Browser (HG16, July2003/NCBI Build 34). Distances of these transcribed regions to the 567 putative promoters were also computed to define matches.

Motif analysis

We examined 400 bp of the 10,567 transcript-matched TAF1 binding sequences (extending 200 bp upstream and downstream from the identified peak) for the occurrence of the TATA box, INR, DPE and BRE elements using matrices defined in TRANSFAC 8.3. and by Chalkley and Verrijzer[130], Kutach and Kadonaga[18], and Lagrange et al[131]. In addition, we examined the conservation of each motif in the promoter regions in chimp, mouse and rat genomes based on a multiple genome alignment [human May 2004 (hg17), chimp Nov. 2003 (panTro1), mouse May 2004 (mm5), rat June 2003 (rn3)]. The motif-matching and conservation scores were calculated using exactly the same algorithm and

cutoffs as described in the UCSC genome browser TFBS conservation track (HG16, July 2003):

http://genome.cse.ucsc.edu/cgi-bin/hgTrackUi?g=tfbsCons

As controls, we checked the randomly selected regions (ENr###s) by ENCODE consortium and generated 12,479 fragments of the same length.

The 10,567 TAF binding sequences (-1000 bp to +200 bp of TAF1 binding site for 5'end-matched and -575 to +575bp of TAF1 binding site for unmatched) and DBTSS promoters (from -1000bp to 200bp of TSS) were also examined for any overlap with roughly 29,000 annotated CpG islands that were documented in the UCSC genome browser [132].

Gene expression analysis

Total RNA from IMR90 cells were extracted using Trizol® reagent (Invitrogen, Carlsbad, CA) and further purified using RNeasy Mini Kit (Qiagen, Valencia, CA) according to manufacturers' recommendations. The purified total RNA was submitted to UCSD Cancer Center Microarray Resource for GeneChip® RNA Expression Analysis using HGU133 Plus 2.0 arrays. The resulting hybridization data was analyzed using Affymetrix GCOS v. 2.0 to determine the detection call as present (P), marginal (M), or absent (A) at significance level $p<0.01$. Detection calls from two technical replicates were combined to give an unambiguous P, M, or A call for each probe set by defining the consensus call as P only if it is P in both experiments, A if it is A in both experiments, and M otherwise.

HGU133 Plus 2.0 probe sets were mapped to corresponding EnsEMBL genes using a translation table downloaded from EnsEMBL (EnsEMBL v26) using the EnsMart

Genome Browser in Nov. 2004. We then evaluated the detection calls of genes only for transcript-matched promoters whose transcripts were aligned to the matching EnsEMBL gene or whose transcripts were matched to the EnsEMBL gene by UCSC Genome Browser tables.

Genes were called P if they were represented by any probe set with a P call and A if all their corresponding probe sets were called A. The set of P or A genes were partitioned by the presence of TAF1-binding to transcript-matched promoters corresponding to the gene in order to define the Class I, Class II, Class III, and Class IV genes.

Multiple promoter usage

To define multiple promoter usage for genes we grouped the set of transcript-matched promoters by the matching EnsEMBL gene and counted the number of transcript-matched promoters for each gene. For this analysis, we considered only those genes for transcript-matched promoters whose transcripts were aligned to the matching EnsEMBL gene or whose transcripts were matched to the EnsEMBL gene by UCSC Genome Browser annotation tables as described above.

Analysis of WEE1 transcripts

Total RNA from IMR90 cells was prepared using Trizol® reagent. cDNA was synthesized from 10 μg of total RNA using poly-dT_{16} primer and Superscript® II reverse transcriptase (Invitrogen). After cDNA synthesis, RNase A/H was added to reaction to hydrolyze RNA, and the remaining cDNA was purified using QIAquick® PCR purification kit, and 20 ng of cDNA was used as templates for quantitative real-time PCR. The primers, 5'-AAGCTGCGACTCTTCGACAC-3' and 5'-

GAGGAGTCTGTCGCACATCA-3' were used to amplify the WEE1 NM_003390 mRNA. The primers, 5'-GAGTACTGCGCAGATGACCA-3' and 5'-GAGGAGTCTGTCGCACATCA-3' were used to amplify the WEE1 AK122837 mRNA. The primers, 5'-GCAAAGACCTGTACGCCAAC-3' and 5'-ACACCGAGTACTTGCGCTCT-3' were used to amplify a reference gene, the gamma actin (ACTG1) mRNA. The quantitative real-time PCR of each sample was performed in triplicate using iCycler and SYBR green iQ SYBR green supermix reagent. The threshold cycle (Ct) values were calculated automatically by the iCycle iQ Real-Time Dectection System Software. Normalized Ct (ΔCt) values for each sample were then calculated by subtracting the Ct value obtained for the ACTG1 gene from the Ct value for the WEE1 transcripts ($\Delta Ct = Ct_{WEE1} - Ct_{ACTG1}$). Using ΔΔCt values (calculated from $\Delta\Delta Ct = \Delta Ct_{WEE1(G0,G1,\text{ or } S)} - \Delta Ct_{WEE1(ASYNCHRONOUS)}$) and the formula, $2^{-(\Delta\Delta Ct)}$, the relative WEE1 transcript levels in cell cycle synchronized sample were determined.

Gene cluster analysis

We sorted the current EnsEMBL genes by chromosome and by annotated start (most 5' end) to define consecutive or neighboring genes. We then searched for runs of consecutive EnsEMBL genes with transcript-matched promoters to define clusters and evaluated the number of runs (clusters) and sizes of runs (clusters). To compute the significance of the number of genes found in clusters of 4 or more consecutive genes (1,668 genes), we randomly selected 6,763 genes in the genome (the number of EnsEMBL genes matched with TAF1-bound promoters) 1000 times and calculated the number of genes found in clusters of 4 or more genes at each iteration (mean=957, standard deviation=59). Given that the observed number of genes in clusters is at least 12

standard deviations away from the mean and no iteration resulted in a count of consecutive genes, in clusters of 4 or more, greater than 1,668, we provide the conservative estimate of significance as $p<0.001$.

Acknowledgments

Chapter 3, is a reprint in full of the material as it appears in: Kim, Tae H; Barrera, Leah O; Zheng, Ming; Qu, Chunxu; Singer, Michael A.; Richmond, Todd A.; Wu, Yingnian; Green, Roland; Ren, Bing. A high-resolution map of active promoters in the human genome. Nature, Vol.436, 2005. I was a primary co-author and researcher of this work. I performed the bulk of the computational portion and analysis of the research. The other primary co-author performed all the experimental assays and validation. Other co-authors of the paper supervised and directed the work or contributed analytical tools (Mpeak) and experimental materials (NimbleGen Arrays).

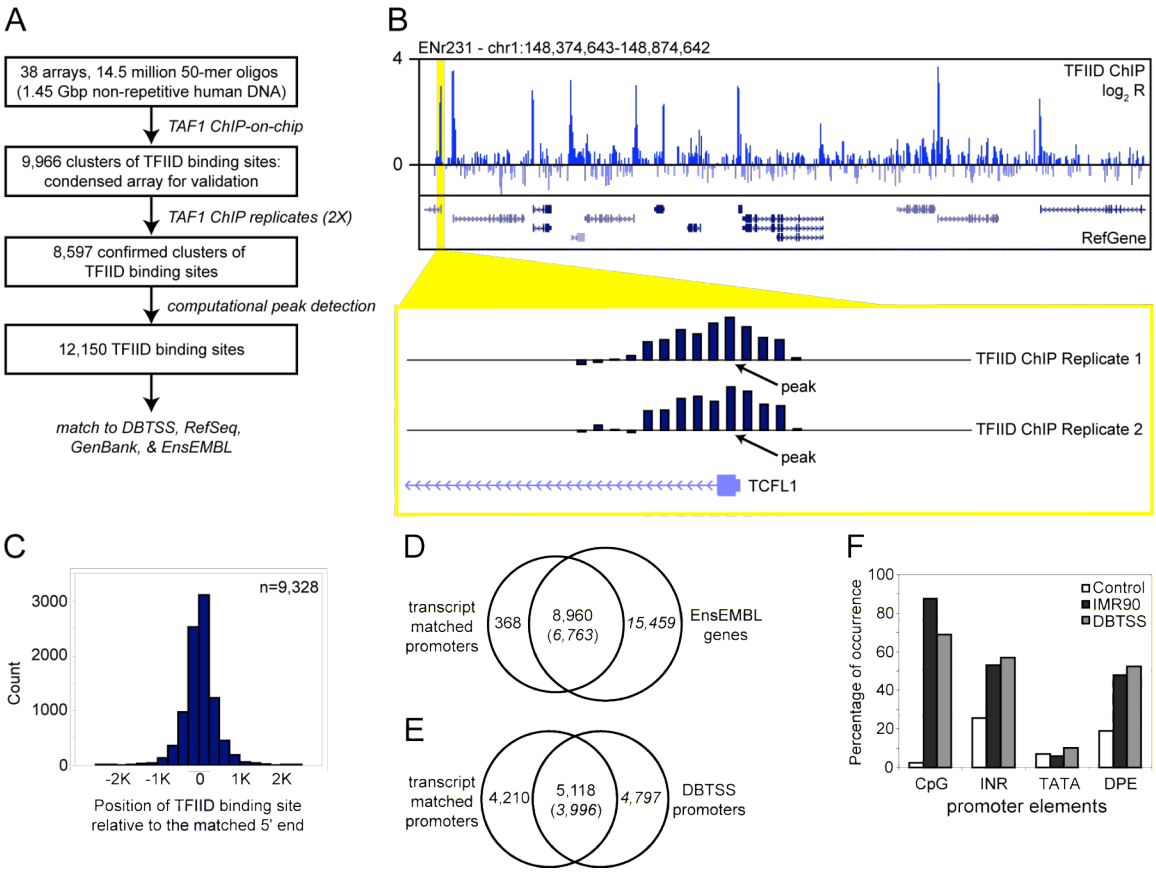


Figure 3-1. Identification and characterization of active promoters in the human genome.

(A) Outline of the strategy employed to map TFIID-binding sites in the genome. (B) A representative view of the results from TFIID ChIP-on-chip analysis. The logarithmic ratio ($\log_2$ R) of hybridization intensities between TFIID ChIP DNA and a control DNA, and RefSeq gene annotation is shown in the top and middle panels, respectively. A close-up view of two replicate sets of TFIID ChIP-chip hybridization signals around the 5' end of the *TCFL1* gene is shown in the bottom panel. Arrows indicate the position of TFIID-binding site determined by a peak-finding algorithm. (B) Distribution of TFIID-binding sites relative to the 5' end of the matched transcripts. (D & E) Venn diagrams showing number of identified promoters that matched EnsEMBL genes (d) or promoters annotated in DBTSS (E). (F) Chart showing the percentage of IMR90 or DBTSS promoters overlapping with CpG islands, or containing conserved TATA box, INR or DPE elements (see Methods for more details).

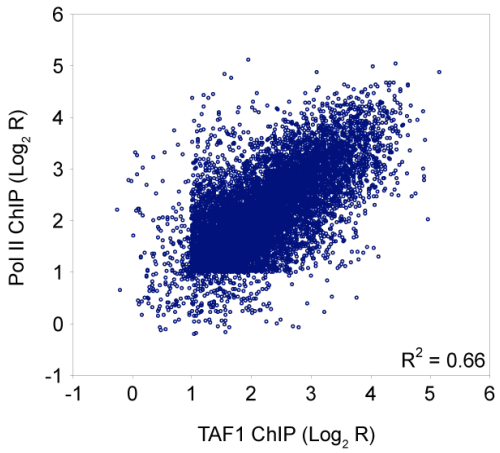


Figure 3-2. TAF1 and Pol II are co-localized.

Hybridization signals (log R) for TAF1 binding sites were plotted against the corresponding signals for Pol II binding. The observed square of the correlation coefficient (R^2) for TAF1 and Pol II was 0.66.

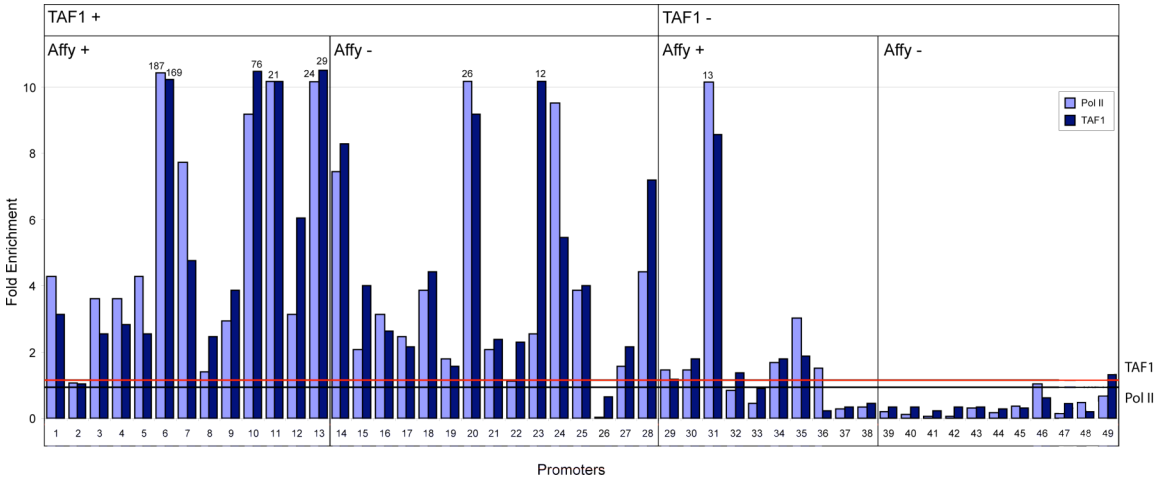


Figure 3-3. Conventional ChIP followed by quantitative PCR validates TAF1 ChIP-chip results.

The x-axis lists all the promoters tested in four classes. The y-axis plots the fold enrichment observed in TAF1 ChIP DNA compared to the unenriched input DNA. The red bar is denotes the value of the average plus two standard deviations observed (1.06) for the all negative controls (class IV) for TAF1. The black bar is denotes the value of the average plus two standard deviations observed (0.94) for the all negative controls (class IV) for Pol II.

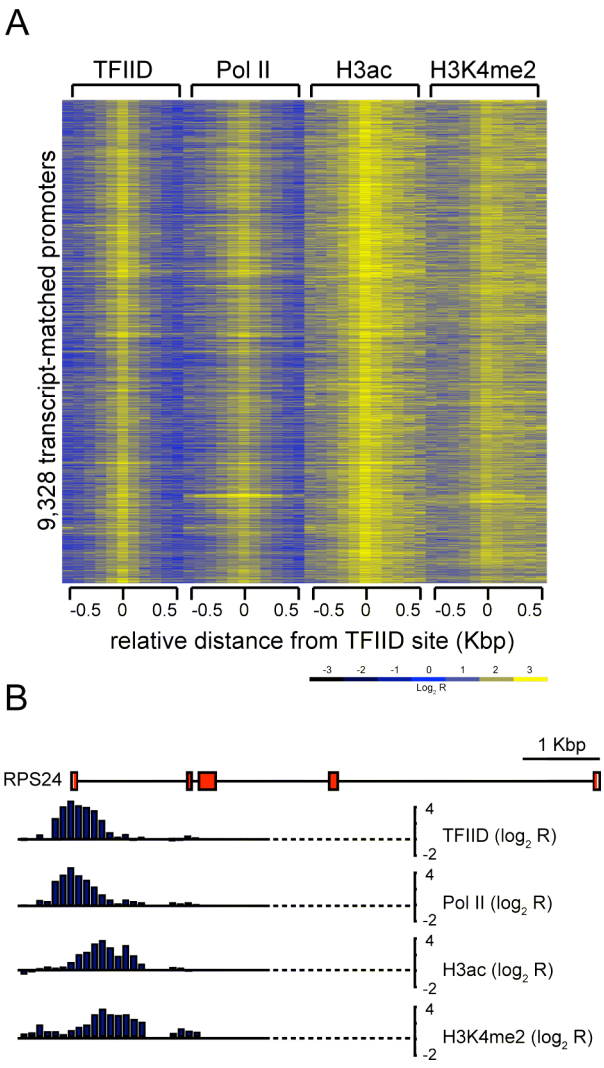


Figure 3-4. Chromatin modification features of active promoters.

Logarithmic ratios of the ChIP-chip hybridization intensities (log_2 R) of probes from 0.5 Kbp upstream to 0.5 Kbp downstream of the identified TFIID-binding sites for TFIID, Pol II, H3ac, and H3K4me2 are plotted in a yellow-blue colored scale for 9,328 transcript-matched promoters. The bottom panel shows a yellow-blue colored scale used to color each cell with corresponding log_2 R values. (b) A detailed view of TFIID, Pol II, H3ac, and H3K4me2 profiles on the promoter of *RPS24* gene.

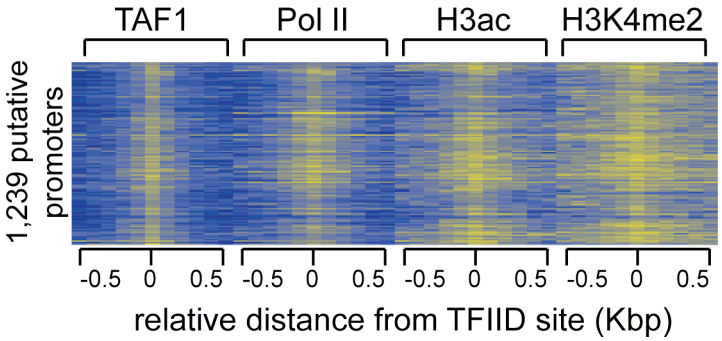


Figure 3-5. Chromatin modifications at putative promoters.

Logarithmic values of the ChIP enrichment ratio (log R) of probes from 0.5 Kbp upstream to 0.5 Kbp downstream of the identified TFIID-binding sites for TFIID, Pol II, H3ac, and H3K4me2 are plotted in a yellow-blue colored scale (the bottom panel of Fig. 2a) for 634 putative promoters.

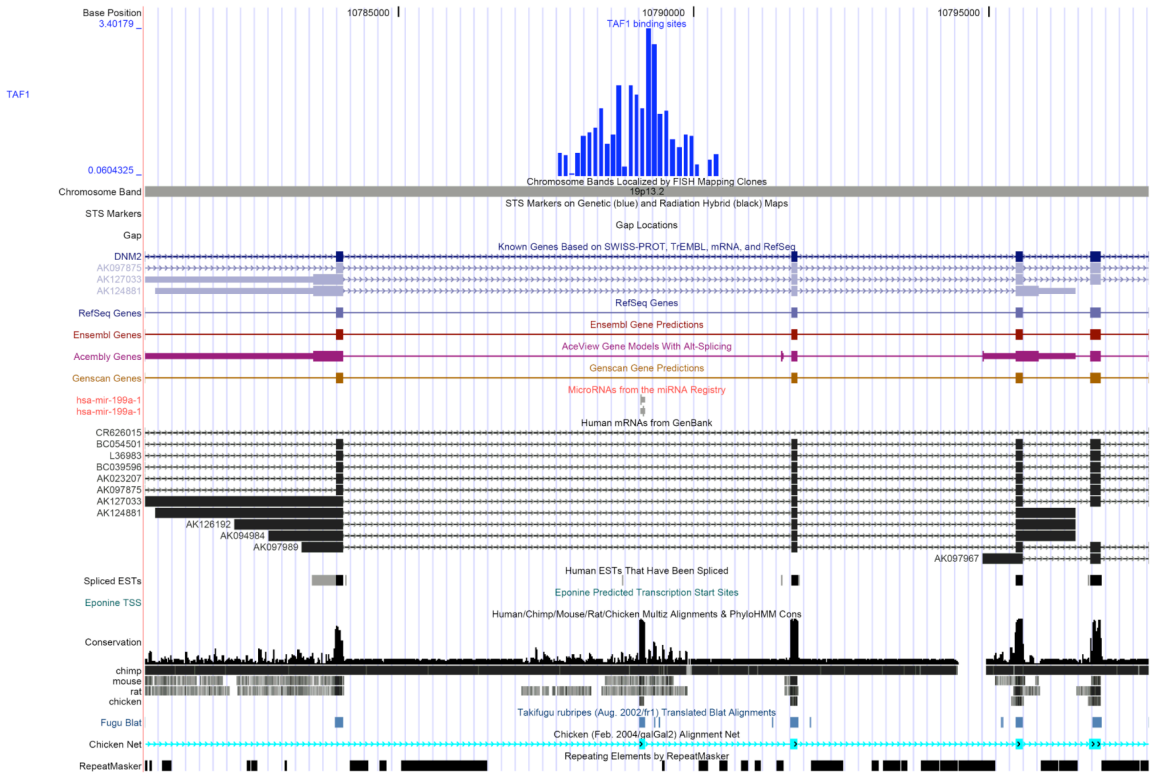


Figure 3-6. A putative promoter maps to a microRNA gene.

UCSC browser window capture shows the TAF1 binding profile and the annotation tracks. The TAF1 binding site is directly over the microRNA has-mir-199a-1.

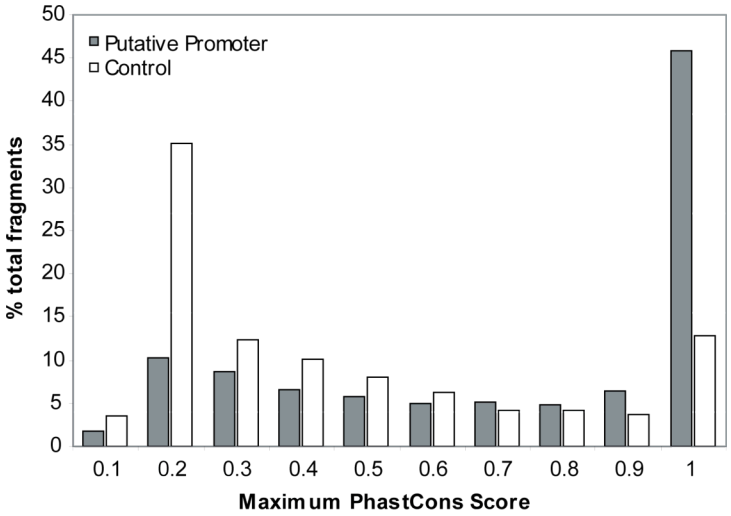


Figure 3-7. Putative promoters are evolutionarily conserved.

Conservation analysis of 1,239 putative promoters (in grey bars) and 1,239 randomly selected control genomic fragments (in white bars) is shown. The x-axis represents conservation score, PhastCon, and the y-axis represents the percentage of all putative promoters (or the control genomic fragments) with the corresponding PhastCon score.

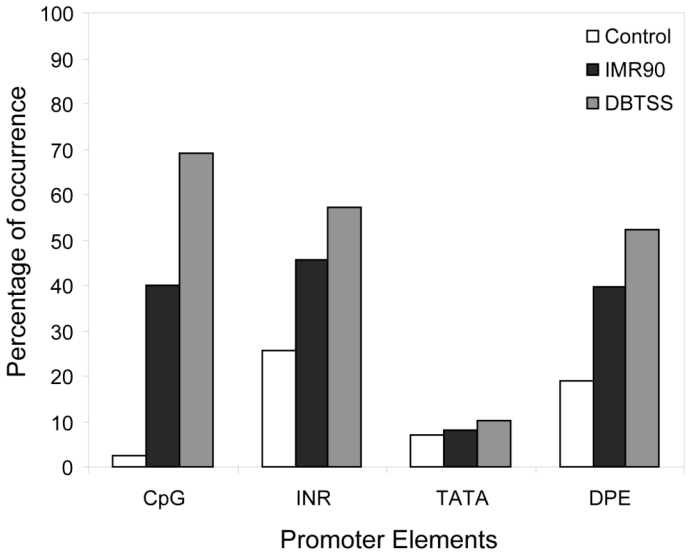


Figure 3-8. Sequence features associated with the putative promoters.

Putative promoters identified in IMR90 cells (referred to as IMR90) were compared to promoters curated in DBTSS or randomly selected genomic fragments. The percentage of putative IMR90 promoters (–250bp to 250bp of TAF1 binding sites) or DBTSS promoters (-1200bp to +200bp of TSS) that overlap with CpG islands, and the percentage of putative IMR90 promoters (from –250 to +250 of TAF1 sites) or DBTSS promoters (-200bp to 200bp of TSS) that contain conserved TATA box, INR and DPE elements are shown. The TATA box consensus is defined by the union of TATAAAT[A/T], [T/A]A[C/T]TTATAT and TTTATA[C/G/T]; the DPE element consensus is defined as [A/G][C/G][A/T][C/T][A/C/G][N]. The INR element consensus is defined as PyPyAN[T/A]PyPy.

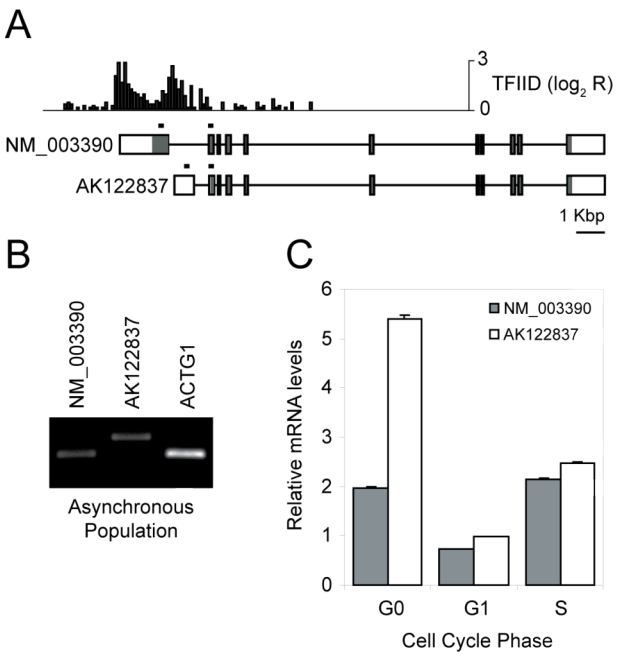


Figure 3-9. Utilization of multiple promoters for a human gene in a single cell type.

(a) Annotation of the *WEE1* gene locus and the corresponding TFIID-binding profile. Black bars over the first and second exons in transcripts indicate the positions of the primers used for real-time quantitative RT-PCR analysis of each transcript. (b) RT-PCR analysis of NM_003390 and AK122837 transcripts in asynchronous population of IMR90 cells. (c) Real-time quantitative RT-PCR analysis of NM_003390 and AK122837 transcripts in cell cycle synchronized population of IMR90 cells. Transcript levels observed for each cell cycle phase were normalized to the level observed in the asynchronous population.

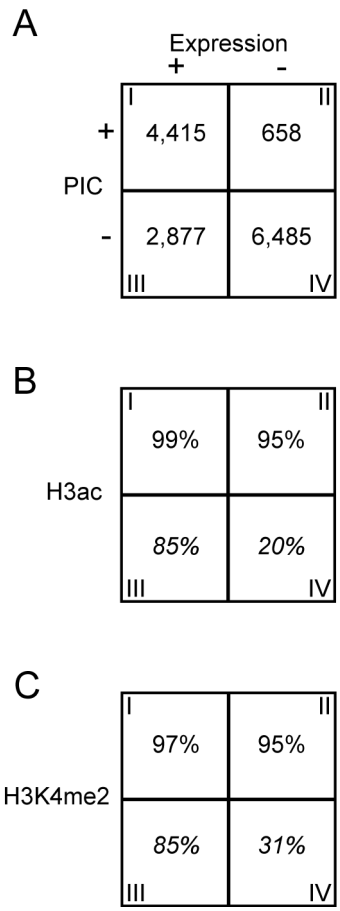


Figure 3-10. Four distinct classes of promoters define the transcriptome of IMR90 cells.

(a) A 2x2 matrix describes the distribution of genes defined by expression and PIC occupancy on the promoter. (b & c) Matrices showing the percentages of genes associated with the H3ac (b) or H3K4me2 (c) modification for each of the four classes of genes. Italicized numbers in some boxes represent extrapolation from the 29 ENCODE regions.

Supplementary Data Tables

The following tables are available for download from:

http://licr-renlab.ucsd.edu/download.html.

They are listed under the publication: Tae Hoon Kim, Leah O. Barrera, Ming Zheng, Chunxu Qu, Michael A. Singer, Todd A. Richmond, Yingnian Wu, Roland D. Green and Bing Ren. A high-resolution map of active promoters in the human genome. *Nature*, 2005.

Data Table 3-1. 9,328 TAF1 binding sites matched to known 5' ends (within 2.5Kbp).

The first column lists coordinates of TAF1 binding sites. The second column lists the database source of the matched transcript. The third column lists the accession numbers for the matched transcripts. The fourth column lists the relative distance between the TAF1 binding site and the 5' end of the matched transcript. The fifth column lists the available gene name for the corresponding transcript. The sixth column lists the corresponding DBTSS ID number. The seventh through 10th columns list ChIP enrichment ratio (log R) of the TAF1, Pol II, H3ac, and H3K4me2 at the TAF1 binding site matched to the transcript. The last four columns list presence of known promoter sequence motifs (CpG, DPE, INR, and TATA respectively) in the promoter.

Data Table 3-2. 1,239 putative promoters (Acembly-match and no 5' end match).

The first column lists coordinates of TAF1 binding sites. The second through fifth columns list ChIP enrichment ratio (log R) of the TAF1, Pol II, H3ac, and H3K4me2 at the TAF1 binding site matched to the transcript. The sixth through ninth columns list presence of known promoter sequence motifs (CpG, DPE, INR, and TATA respectively). The 10th column lists the matched Acembly gene models. The 11th column lists IDs for those EnsEMBL genes that contained TAF1 binding sites within the gene locus, but not at their 5' ends.

Data Table 3-3. Validation of putative promoters.

The 35 novel transcribed units identified within 2.5Kbp of the putative TAF1 binding sites are isolated in the worksheet "Novel transcribed units." The first column lists coordinates for transcriptionally active regions detected in validation array experiment. The second column lists transcriptionally active cluster ID. The third column lists the coordinates for the putative promoters identified by TAF1 genome-scan that are within 2.5 Kbp of the detected transcript. The fourth column lists the accession number of known transcripts whose exons overlap with the detected transcript. The fifth column lists the accession number of those transcripts whose exons flank the putative promoter and detected transcript. The last column lists the coordinates for the closest putative promoters identified by TAF1 genome-scan from the detected transcript.

Data Table 3-4. Multiple promoter usage.

The first column lists the EnsEMBL genes with 2 or more TAF1 binding sites. The second column lists gene names. The third column lists the number of TAF1 binding sites found with the gene. The fourth column lists the accession numbers of matching transcripts and the coordinates of the corresponding TAF1 binding sites.

Data Table 3-5. Clusters of genes with TAF1-bound promoters.

The first column lists the cluster size in number of genes. The second column lists EnsEMBL gene IDs of the genes in the cluster. The third column lists the gene names.

Data Table 3-6. Gene expression classes.

Four classes of genes as discussed in the text are appended in order. The first column lists the gene names. The second column lists EnsEMBL gene IDs.

Data Table 3-7. Histone modification on class III and IV genes.

Two separate worksheets, "ClassIII_ENCODE" and "ClassIV_ENCODE" are provided. The first, second and third columns list the chromosome ID, start and end coordinates respectively of the EnsEMBL gene found within the 29 ENCODE regions analyzed. The fourth column lists EnsEMBL gene IDs. The fifth column lists gene names. The sixth column lists the ENCODE region IDs. The seventh through 10^{th} column lists whether TAF1, Pol II, H3ac and H3K4me2 are bound at the corresponding promoter, respectively. The last column provides additional notes.

Chapter 4

Genome-wide Mapping of Tissue-specific Promoters

The analysis of several mammalian genomes has revealed between 20,000 to 30,000 genes in each genome, a number that may seem hard to reconcile with the large number of cell types and complex functions of these organisms. The solution to this paradox partly lies in the large array of transcripts that each gene can potentially generate through usage of alternative promoters and the variable levels of transcripts that each gene produces in different tissues and cell types. Thus, in order to understand the mechanisms that control diverse patterns of gene expression in mammals, it is necessary to accurately define the active promoters and monitor their cell or tissue-dependent activity. Previous high throughput strategies for assaying tissue-specific gene expression have primarily relied on measurements of steady-state transcript levels by microarrays or tag sequencing. Here, we employ a new experimental strategy to identify and characterize tissue specific promoters by integrating genome-wide maps of RNA polymerase II (Pol II) binding, chromatin modifications and gene expression profiles. We applied this strategy to mouse embryonic stem cells (mES), and adult brain, heart, kidney, and liver. Our results delineated 24,363 Pol II binding sites throughout the genome, 91% of which correspond to 5' end annotation based on known transcripts and cap-analysis of gene expression (CAGE) and can be regarded as promoters. A majority of these experimentally defined promoters are active in all tissues, while only 4,396 can be characterized as tissue-specific using a quantitative measure of Pol II occupancy. In

general, Pol II occupancy at these tissue specific promoters is correlated with the presence of active histone modification marks. However, a set of mES- specific promoters display persistent levels of H3K4me3 in non-ES tissues despite undetectable Pol II binding and transcript. Broadly, our results expand the knowledge of tissue-specific mammalian genes and provide a resource for understanding the transcriptional programs in mammalian development and differentiation.

4.1 Introduction

Mammalian organisms are characterized by a diversity of cell types -- from the zygote and progenitor cells to the more than 200 differentiated cell types which perform the functions of organs in adults. In general, the functions of each cell type are specified by the complement of genes expressed [28,58,119,133,134]. Thus, investigating the mechanisms that control gene expression, beginning with the critical step of transcription initiation, will contribute toward understanding how the diversity of mammalian cell types and their functions are generated [135].

Many large-scale efforts have been devoted to the investigation of transcript expression patterns across cell and tissue types. Microarray-based technologies and high throughput sequencing methods have been used to determine steady-state mRNA levels of genes in a compendium of cell and tissue types under normal or pathological conditions [28,119,133,134,136]. In addition, recent advances in the sequencing of transcript 5' ends have also expanded the annotation of mammalian promoters in different mammalian tissues and provided valuable references of potential transcriptional start sites for most mammalian genes [29,137,138]. These studies have revealed a large spectrum of transcripts

for each gene generated by extensive usage of alternative promoters, alternative splicing and alternative polyadenylation sites. One of the questions raised by these observations is how cells control the usage of different promoters to produce the diverse forms of transcripts.

In order to understand the mechanisms that drive differential gene expression in diverse cell and tissue types, it is necessary to examine transcription factor binding and chromatin structures at the active promoters for each gene in various tissues. Transcription of protein-coding and most non-coding RNA genes starts with the binding of RNA polymerase II (Pol II) over the core promoter spanning the transcript start site (TSS) [12,19]. In eukaryotes, transcriptionally active promoters have been associated with nucleosome depletion at some TSS, as well as characteristic histone variants and histone tail modifications (H3K4me2, H3K4me3, and H3ac) [38,45,50,58-60,139,140]. In addition, binding sites for sequence-specific transcription factors have been identified within promoters and linked to cell or tissue-specific patterns of expression [24,25,141-146].

In this study, we employ an integrated experimental strategy to characterize transcriptional promoters in the mouse genome across a panel of mouse organs – brain, heart, kidney, liver – and mouse embryonic stem cells (mES). We first identify active promoters by localizing the binding sites of the RNA polymerase II pre-initiation complex in each tissue throughout the genome. We then confirm the activity of the promoters by examining the active chromatin marks including histone H3 acetylation and methylation. We assess promoter tissue-specificity by promoter Pol II binding, active chromatin modifications (H3ac, H3K4me3) and relative transcript levels.

Our results lead to the identification of 24,363 Pol II binding sites which match existing transcript annotation as well as 5' end sequencing from cDNA libraries of various mouse tissue and cell types [93,137,138,147,148]. By adapting a definition of tissue-specificity based on Shannon entropy previously used for gene expression data, we define 4,396 promoters as having enriched Pol II binding in a particular tissue – adding to the annotation of tissue specific promoters in the literature [27]. Microarray analysis of gene expression across tissues supports the classification of genes with tissue-specific Pol II enrichment. The combination of these data leads to a high-confidence catalog of genes which contribute to the uniqueness of the tissues surveyed. Additionally, we identify known and novel sequence motifs that characterize tissue-specific promoters in brain, heart, kidney, liver, and mES. Epigenetic patterns of acetylation and lysine 4 tri-methylation of histone H3 defined by ChIP-chip generally correspond with Pol II enrichment across tissues. However, half of the promoters with enriched Pol II binding in mES maintain active epigenetic marks in tissues where Pol II binding is relatively depleted. Broadly, these findings underscore the utility of Pol II binding and chromatin modification data as key resources complementing transcription profiling in unraveling the layers of tissue-specific gene regulation.

4.2 Genome-wide Mapping of Pol II in mouse ES cells and adult tissues

4.2.1 Overview of Strategy

In eukaryotes, RNA polymerase II (Pol II) drives the synthesis of mRNA and small nuclear RNA. Its C-terminal domain (CTD) is hypo-phosphorylated as part of the pre-initiation complex (PIC), phosphorylated at Ser5 early in the transcription cycle and

at Ser2 toward the 3' end of the gene [149,150]. Thus, we used a monoclonal antibody (8WG16) specific for the hypo-phosphorylated RNA polymerase II CTD to map PIC binding at active promoters in mouse brain, heart, kidney, and liver tissue, as well as R1 ES cells using chromatin immunoprecipitation with microarrays. We adapted the strategy we previously used to map active promoters in human fibroblast cells (Figure 4-1) [43]. Specifically, we first performed ChIP-chip on Pol II using chromatin prepared from the four organs and ES cells, and a set of 37 microarrays, containing a total of 14.3 million 50-mer oligonucleotides, tiling the non-repetitive sequence of the mouse genome at 100 base-pair (bp) resolution. The results from the genome-wide survey of Pol II binding led to the identification of a total of 32,482 Pol II binding sites. We designed a set of four microarrays containing 1.4 million oligonucleotides to cover each site extended by 2 Kbp upstream and downstream, and repeated independent Pol II ChIP-chip for each tissue to confirm Pol II binding (condensed scan). To define confirmed sites of Pol II binding, we applied our previously described peak finding strategy on the condensed scan ChIP-chip and genome-scan ChIP-chip for each tissue [43,72]. We required that a peak of Pol II binding predicted in the condensed scan is within 500bp of a peak predicted in the genome scan (Figure 4-1, Methods).

4.2.2 Summary of Pol II Binding and Annotation

Using the procedure summarized in Figure 4-1, we defined a total of 24,363 high-confidence, non-overlapping Pol II binding sites in the mouse genome across five tissues (Data Table 3-1). Each of these sites has confirmed binding based on the genome scan and condensed scan for at least one tissue. These binding sites range in size from 50bp to 18Kbp. By assaying Pol II enrichment by ChIP with quantitative PCR (ChIP-qPCR) at

27 randomly selected gene promoters in mES cells, we estimated ~ 70% sensitivity and 100% specificity for our method of defining Pol II binding sites by ChIP-chip in each tissue (Figure 4-2). Additionally, we estimated a 100% positive predictive value (PPV) by ChIP-qPCR validation of 24 randomly selected Pol II ChIP-chip bound sites in liver (Figure 4-3).

Since the PIC-form of Pol II is expected to localize over transcription initiation sites in the genome [150-152], we compared the location of these binding regions with annotated mRNA transcript start sites (TSS) downloaded from the UCSC Genome Browser (MM5; refGene, knownGene, ensGene, and all_mrna) [93]. 16,976 (69%) of these sites mapped within 2.5Kbp of (66,559) distinct transcript start sites (TSS) based on RefSeq, Ensembl, UCSC knownGene, or GenBank annotation. These transcripts in turn correspond to 11,000 out of ~24,000 mouse genes based on Entrez Gene annotation [94]. Of the remaining unmatched sites within and outside of known gene loci, 5,153 mapped within 2.5Kbp of TSS based on 5'cap-analysis of gene-expression (CAGE) sequencing from a panel of 145 mouse cDNA libraries [138,148]. Taken together, these two lines of evidence provide independent support that 91% of these Pol II binding regions correspond to known transcription initiation sites (Table 4-1).

The distance distribution of Pol II binding sites to matching TSS clearly supports the accuracy of our method in defining known transcription initiation sites (Figure 4-4). In addition, the number of promoters relative to the number of genes suggests the prevalence of alternative promoter usage. For instance, a recent RNA interference study defined estrogen receptor beta (*Esrrb*) as one of 7 genes which are critical for embryonic stem cell renewal in vitro [153]. We identified two tissue-specific promoters for this gene;

one of them appears to have enriched Pol II binding in mES, while the other shows enriched binding in kidney (Figure 4-5). We estimate that 28% of genes with Pol II binding utilize two or more alternative promoters across the five tissues. This estimate is half of the previous estimate in mammalian genomes and may be due to the limited number of tissues surveyed as well as the more limited resolution of transcription initiation sites based on Pol II binding compared with the base-pair resolution of 5' end sequencing methods [29,137].

Additionally, in characterizing the genomic distribution of the CAGE-matched sites, we validate estimates of exonic transcription initiation activity based on CAGE data [137]. The majority (62%) of the CAGE-matched sites resides within known gene boundaries (exonic and intronic) (Figure 4-4). A substantial fraction are tissue-specific (37%) and the prevalence of these sites underscores the role of transcription initiation, along with splicing, in defining the complexity of transcript populations even from within known gene loci. A previous study based on CAGE tag frequency has correlated this exonic promoter activity with tissue-specific genes [137].

4.2.3 Novel Promoters

By examining the co-localization of H3K4me3, an epigenetic mark associated with 5' ends of active genes from yeast to human, we defined 382 sites not near known TSS or CAGE tag clusters as putative promoters. This fraction (1.6%) of our catalog suggests only a small number of transcription initiation sites still missed by extensive 5'end sequencing efforts to annotate the mouse transcriptome (Figure 4-4). A large fraction (37%) of these putative promoters appears to be tissue-specific. These putative

promoters are primarily from mES (67%) and kidney (18%). Further investigations are necessary to determine the matching transcripts for these uncharacterized promoters.

4.2.4 Unexpected Pol II Binding Behavior

We also observed examples of unusual patterns of hypo-phosphorylated Pol II binding deviating from its expected localization over canonical TSS [151,152]. For instance, distinct binding sites for Pol II were found within boundaries of known transcripts in addition to the TSS (Figure 4-6). Of 3,843 genes surveyed with multiple Pol II binding sites within transcript boundaries, 46% have significant correlation between increased Pol II binding density and relative tissue expression (R>0.6, 2-fold enrichment above expectation). Furthermore, contrary to expected punctuate patterns of Pol II binding at TSS, we catalogued 53 broad tissue-specific Pol II binding regions greater than 5Kbp in span (Figure 4-7, Data Table 4-2). 44 of 53 overlap transcript start sites for known genes. These overlap highly expressed tissue-specific genes such as the cardiac muscle protein leimodin (*Lmod2*) and a secreted protein, natriuretic peptide precursor A, (*Nppa*) in heart. Additional examples of these regions overlap well-known ES-cell enriched genes such as *Pou5f1* and *Rcor2* (Figure 4-7 D). The Pol II binding patterns for these two genes extend thousands of base pairs upstream as well as downstream from the TSS (Figure 4-7A,B). H3ac mimics the Pol II binding pattern while H3K4me3 appears concentrated over the TSS in comparison. Finally, among 819 tissue-specific Pol II binding sites not matched to known 5' end data based on annotated transcripts or CAGE or H3K4me3 localization, we found approximately half to be enriched in mES cells (Table 4-2). A significant fraction of these mES sites overlap previously mapped Oct4 and Nanog binding sites in

mES (24%) compared to expected (0.2%) [154]. Although our annotation clearly links the majority (92%) of the Pol II bound to transcription initiation, this preliminary analysis suggests that a small fraction may be linked to distal regulatory sites which physically interact with promoters [155,156].

4.3 Tissue-specific Promoters

4.3.1 Entropy Measure of Tissue-specificity

In order to characterize the tissue activity of a particular promoter or Pol II binding site, we used the ChIP-chip $\log_2$ratio enrichment as a measure of Pol II occupancy at all sites across tissues and defined an index of tissue activity for each site by adapting a Shannon entropy previously applied to microarray gene expression and EST data [27]. We defined the relative Pol II binding in a tissue *t* for a given site *s* as $p_{t|s} = B_{t,s} \Big/ \sum_{1 \le t \le N} B_{t,s}$ where $B_{t,s}$ is the average ChIP-chip $\log_2$ratio in the 1Kbp neighborhood centered at the midpoint of Pol II binding site *s,* and *N* is the total number of tissues surveyed. The entropy of a site's Pol II binding distribution across tissues is then defined as: $H_s = \sum_{1 \le t \le N} p_{t|s} \log_2 p_{t|s}$. The measure H_s has units of bits and as in its use with expression data, the value of H_s ranges from zero, for genes bound by Pol II in a single tissue, to $\log_2(N)$ for sites bound uniformly in all tissues surveyed. We also adapted the companion measure of "categorical tissue-specificity" to characterize the bias of a Pol II binding site for a particular tissue defined as $Q_{s|t} = H_s - \log_2(p_{t|s})$. This index also has units of bits and as before has a minimum of zero when a site is bound by Pol II

predominantly in the tissue and grows without bound as the relative binding of Pol II in that tissue goes to zero.

4.3.2 Tissue-specific MicroRNAs

We used these measures of entropy and categorical tissue-specificity to assess the usage of all Pol II binding site across tissues. When applied to sites not matched to known mRNAs but near known microRNAs (miRNAs), 10 of 19 matched miRNAs were classified as tissue-specific. Recent studies have provided evidence that miRNAs play a pivotal role in defining tissue and cell-specific expression patterns (Table 4-3) [157]. Indeed, 7 of the 10 promoters we defined as tissue-specific for the miRNA were cloned from the corresponding tissue source, or closely-related tissue source in the case of mES and testis [158]. Two of these tissue-specific miRNAs have been shown to downregulate a large number of miRNAs in human: miR-124 transfection in HeLa cells shifted the expression profile towards that of brain, while miR-1 shifted the expression profile of HeLa cells toward heart and skeletal muscles [157].

4.3.3 Promoter Tissue-specificity and CpG Islands

Overall, the majority of transcript-matched promoters have ubiquitous activity by the Pol II binding entropy across the tissues surveyed (Figure 4-8). As expected, the promoters uniformly bound by Pol II overlap significantly with CpG islands compared to promoters with Pol II binding enriched in specific tissues [27,137,159,160]. Tissue-specific promoters defined by a low entropy measure ($H \leq 1$) have a five-fold decrease in CpG island overlap (15%) compared with promoters with a high entropy measure ($H \geq 2$) associated with ubiquitous activity (75%). Profiling of Pol II and active chromatin

modifications at CpG versus non-CpG island promoters suggests that nearly all promoters overlapping CpG islands have some H3K4me3 across tissues even when Pol II binding and H3ac appears weak (Figure 4-9). ChIP-qPCR of Pol II and H3K4me3 enrichment at 5 randomly selected promoters with variable Pol II occupancy supports this observation (Figure 4-10). Subtle enrichments of H3ac and H3K4me3 revealed by these promoter profiles across tissues are not likely to be called "present" by typical ChIP-chip analysis methods and reveal the limitations of binary calls in ChIP-chip analysis.

4.4 Tissue-specific Gene Promoters

To hone in on the relationships among promoter Pol II binding, active chromatin modifications, and transcript level in tightly regulated expression, we focused the remainder of our analysis on 9% of the gene promoters (937) with Pol II binding enriched in a specific tissue and profiled the Pol II, H3ac, and H3K4me3 ChIP-chip $\log_2$ratios 2Kbp upstream and downstream from a reference start site and compared the matching gene expression enrichment by normalized expression signal across tissues (Figure 4-11).

4.4.1 Promoter Pol II Binding and Expression

The panels illustrate that tissue-enriched Pol II binding correlate as expected with higher gene expression levels in that tissue relative to other tissues, not just based on our expression array data but also from a compendium of expression data from 61 mouse tissues [28,119]. To quantitatively measure this correlation, we created ranked lists of all genes for each tissue ordered by their categorical tissue-specificity based on our expression data [27]. We then assessed the enrichment of each set of genes defined as

tissue-specific based on Pol II binding at the top of the ranked list for each tissue based on categorical tissue-specific expression. Not surprisingly, the measures of categorical tissue-specificity using binding and expression data correlate significantly (Table 4-4). We highlight the top ten tissue-specific genes defined by expression within each set of genes defined as tissue-specific based on Pol II binding (Figure 4-12). Among these genes are those known to be highly-specific and highly-expressed in heart such as cardiac myosin (*Myl2*) and actin (*Actc1*) as well as mES-enriched genes reported to be characteristic of stem cells such as *Tdgf1, Zfp42, Nanog, and Pou5f1*.

Comparison of genes defined as tissue-specific based on binding and expression allows the identification of a high-confidence set of genes with tissue-enriched activity. Conversely, examining the genes defined as tissue-specific by Pol II binding not supported by expression data can be useful in identifying possible mis-assignment of Pol II binding to a gene based on the nearest 5' end assumption or the transcript to gene mapping annotation. Alternatively, this minority might represent tissue-specific promoters for genes which might be regulated at steps beyond initiation [149]. For instance, two genes with enriched Pol II binding and histone modifications at its promoter region have no enrichment in mES based on our expression profiling data. *4930511H11Rik* appears to be more highly expressed, albeit in low levels in adult tissues, while *Tmcc3* is called absent across the tissues we surveyed. Based on the GNF expression atlas, *4930511H11Rik* appears to be selectively expressed in testis, while *Tmcc3* is selectively expressed in the oocyte and fertilized egg (Figure 4-13).

4.4.2 Pol II Binding and Histone Modifications

Across tissues, tissue-specific Pol II enrichment matches enrichment of epigenetic marks generally associated with gene activity (Figure 4-11). In mES, however, genes with specific Pol II enrichment can be further partitioned into two major classes. The first category (mES c1) suggests a "strict" mechanism for defining cell-specific transcription initiation and expression similar to the general profile observed among promoters with enriched activity in specific adult tissues. For example, Pol II and histone modifications are enriched only in mES and not detectable by ChIP-chip in other tissues as shown for the *Lin28* gene (Figure 4-14 A). The second category (mES c2) shows that although there appears to be preferential gene expression enrichment and Pol II binding in mES, other tissues have clearly detectable although weaker histone modifications over the promoter region of the same gene as exemplified by the Pol II and H3K4me3 promoter profile and gene expression of *Dnmt3b* across tissues (Figure 4-14 B).

ChIP with quantitative PCR (qPCR) for Pol II, H3K4me3, and H3ac at four genes from each mES category confirm the Pol II enrichment at these promoters specific to mES. We also verify the partitioning of these two categories by the relative enrichment of histone modifications, in particular of H3K4me3, in adult tissues for mES c2 (Figure 4-15). Pol II binding enrichment is at least 5-fold greater in mES compared to all other tissues for each gene promoter in both c1 and c2 (Figure 4-15 A). Relative enrichment of H3ac in adult tissues for promoters in c2 appears lower than in mES, but this detection in adult tissues is notable relative to promoters in c1 and the control (Figure 4-15 B). H3K4me3 enrichment appears be comparable between adult tissues and mES at mES c2 promoters with the exception of the *Sox2* promoter (Figure 4-15 C). For *Sox2*, a minor

H3K4me3 enrichment was observed in brain only, and clearly less than in mES. Although Sox2 has been implicated in embryonic stem cell self-renewal and pluripotency in concert with Nanog and Oct4 [63], it is also found to be present as a neural stem cell marker with suggested roles in neuron maintenance in the adult brain [161].

4.4.3 Functional Characterization of Tissue-specific Genes

To compare our grouping of genes based on tissue-enriched Pol II promoter binding with existing functional annotation, we determined the enriched GO biological process (GO-BP) categories in each group [136,162]. We found that the most enriched GO-BP categories correspond to the known physiological roles of the tissue and cell type (Table 4-5).

Here, we highlight genes from each set with known regulatory roles as well as groups of genes whose known biological functions characterize the tissue. For instance, among the brain-enriched genes are myelin transcription factors (*Myt1, Myt1l*), homeobox proteins (*Pknox2, Uncx4.1*), zinc-finger proteins (*Egr3, Scrt1*), and forkhead factors (Foxg1). We also recovered a number of genes involved in synaptic transmission such as acetylcholinesterase (*Ache*), GABA receptors (*Gabra1, Gabrg2*), and glutamate receptors (*Gria2, Grin2b*). In heart, we recover developmental regulators such as *Hand2, Nkx2-5, Smyd1, Sox6, Tbx18,* and *Tbx20* in addition to genes for cardiac muscle proteins such as *Actc1, Mybpc3, Myh6, Myl3, Myom1, Tnnc1,* and Ttn. We also found several Hox genes to have enriched expression in kidney (*Hoxa7, Hoxa9, Hoxa10, Hoxb6, Hoxb9, Hoxc6, Hoxc10, Hoxd3, Hoxd9,* and *Hoxd10*). These pattern specification genes have been suggested to be critical for renal organogenesis [163]. In liver, we find the known

hepatic nuclear factor 3 (HNF3)/Forkhead family transcription factors *Foxa1, Foxa3* as well as the related HNF6 transcription factor *Onecut1* [57]. Several of the aforementioned developmental regulators found to be enriched in adult tissues are repressed by Polycomb group proteins in mouse embryonic stem cells [164] such as *Egr3* and *Uncx4.1* in brain [165], key heart developmental genes such as *Nkx2-5*, *Tbx18*, and *Tbx20* [166,167], *Pax8* and Hox genes in kidney [168], and the liver factor *Onecut1*.

In mES, we observe that the two classes of gene promoters have a subtle difference in the ranking of the most enriched GO-BP categories. The mES c2 class is most enriched in genes related to cell cycle and cell division, while mES c1 is most enriched in genes related to cell proliferation and pattern specification. Among the genes in mES c2 are those which may not have restricted expression in mES but clearly enriched activity such as a host of cell-cycle related genes (*Ube2c, Sgol2, Bub1, Bub1b, Aurkb, Cdc2a, Cdca2, Cdca7, Cdc25c*) and DNA replication genes (*Mcm3, Mcm8*). Among genes in mES c2 with reported roles in development are Gli zinc-finger transcription factors (*Gli1, Gli2, Zic3*) activated through the Sonic hedgehog (Shh) signal-transduction pathway as well as a hedgehog receptor gene, *Ptch2* [169]. *Gli1* and *Gli2* — both of which mediate Hh signals — have been implicated in tumorigenesis and are reported to be found among precursor cells in adult tissues [169]. Additionally, the lymphoid enhancer factor 1 (*Lef1*) gene, which mediates the effects of the Wnt signaling pathway, belongs in this class [170].

Among the mES c1 genes, we find the majority of genes which have studied roles in stem-cell renewal and pluripotency such as *Pou5f1, Nanog* [63,154], as well as additional stem-cell markers such as *Dppa4, Nr0b1, Utf1, Tdgf1, Zfp42* [171,172]. We also define

previously identified ES-enriched genes in the TGF-beta signaling pathway such as *Lefty1, Lefty2,* and *Nodal* [171,173] as well as fibroblast growth factors such as *Fgf4, Fgf15*, and *Fgf17*. Among these FGFs, *Fgf4* has a reported role in trophoblast stem cell proliferation [174]. Because the comparison of Pol II binding in mES is relative to adult tissues, genes with reported roles in development were also found in mES c1. These may not necessarily be ES-specific transcription factors, but may have poised promoters marked by Pol II binding and H3K4me3 or basal transcriptional activity. *Gbx2* has reported roles in nervous system development [175]; *Pitx2*, heart development [176]; *Six6os*, eye development [177].

4.4.4 Sequence Motifs at Tissue-specific Promoters

Nearly half (45%) of the promoters in mES c2 overlap CpG islands. This proportion is more than two-fold higher than the overlap of promoters in mES c1 with CpG islands (20%). Among the adult tissues, brain appears to have the largest overlap (24%) between tissue-specific gene promoters and CpG islands compared with heart (10%), kidney (14%), and liver (9%). This is in agreement with a previous observation that among transcripts with specific expression patterns, promoters associated with the central nervous system were exceptionally CpG-rich [137].

In order to define discriminating sequence motifs within each tissue-specific promoter set, we use two complementary motif-finding strategies. The first strategy measures motif enrichment in each tissue-promoter set relative to a background set based on a balanced error measure which equally weighs a motif's ability to identify promoters in the set (sensitivity) and to correctly discriminate against promoters not in the set

(specificity) [25,143,144]. Using this strategy, we characterized the enrichment of known vertebrate motifs from TRANSFAC [178] and JASPAR [179] in each tissue-specific promoter set relative to two types of background promoter sets: (1) a random set of mouse promoters from CSHLMPD [142], and (2) the relative complement of the tissue-specific promoter set in the set of all tissue-specific promoters (Table 4-6). To identify novel motifs in each tissue-specific promoter set, we used a previously described *de novo* motif finder, DME [25,143,144]. We evaluated the significance of these novel motifs using the same misclassification metric and report the novel motifs for each set (Table 4-6).

As a complement to this strategy, we used relative over-representation of conserved occurrences to define characteristic motifs for each tissue set. Strictly defining a conserved occurrence as the best match to the motif aligned in the same position at orthologous mouse and human promoters, we identified binding sites for transcription factors with previously reported roles in the specific tissue or cell type, as well as others whose roles remain unclear or whose binding domains appear similar to those of transcription factors with reported roles in that tissue (Table 4-6).

4.5 Discussion

One of the first steps towards a comprehensive understanding of the mechanisms of cell diversity is to define and profile the active promoters in different cell types. Here we described an integrated approach for profiling the epigenetic and sequence features of active promoters in mouse embryonic stem cells and four adult organs defined by genomic Pol II binding. We defined 24,363 Pol II binding sites that correspond to complementary evidence of transcription initiation based on known transcript 5' ends and

CAGE annotation (91%). Our study provided evidence for over 5,000 TSS previously supported by CAGE evidence alone, confirmed widespread usage of alternative promoters by mammalian genes, and identified several thousand promoters as tissue-specific. These tissue specific promoters led to the identification of transcription factor motifs potentially related to tissue specific transcription factor binding, genes with tissue specific expression, and a class of ES cell genes with promoters persistently marked by active chromatin modifications in adult tissues.

To characterize the tissue-specificity of factor binding by ChIP-chip at promoters, we adapted a quantitative index based on Shannon entropy. This strategy overcomes some of the limitations associated with ChIP-chip technology. The current emphasis on "bound" versus "unbound" sites in ChIP-chip analysis sacrifices sensitivity for specificity in defining sites associated with a particular factor. This naïve classification becomes especially problematic, however, when comparing factor occupancy at genomic sites across cell types or conditions. Further development of quantitative measures of relative ChIP-enrichment for a factor's genomic localization across samples or conditions, as used here, will be critical in circumventing these issues.

We used two complementary approaches – classification and conservation – to define the sequence motifs associated with tissue specific promoters based on our entropy measure. We recovered binding sites for the "master regulator" HNF4 as a significant known motif in liver, and a binding site for the muscle regulator SRF as a significant motif in heart [25,65,143,144,180]. A Myc-Max binding motif, enriched in the mES c2, supports a purported key role for c-myc in ES cell regulation [181]. Several motifs for CREB and its related factors such as ATF underscore their widespread roles in the brain – in memory

formation, neuronal plasticity and survival [182]. Although it has been mainly associated with MHC class II gene transcription, we also found the Rfx5 binding motif to be enriched in our brain-specific genes. A unique DNA binding domain characterizes the RFX family of transcription factors and the roles of other family members are uncharacterized in mammals. A previous study has shown that the Drosophila homologs of the mammalian *Rfx1* to *3* genes are detected only in the embryonic brain and peripheral nervous system of the fly [183]. Additionally, Ap4, Mef2, and specific muscle TATA-binding protein (TBP) motifs, previously linked to muscle-specific expression were identified in heart along with motifs for orphan nuclear receptors such as Rora and Sf1. Sf1 has no clear role in the heart or muscle cells based on literature search, while Rora has been implicated in the regulation of genes involved in lipid homeostasis of skeletal muscle [184]. The Rora motif in combination with a Tcf11 motif has also been shown to be enriched at promoters of heart-specific genes [185]. In kidney, a binding motif for a key regulator of renal development, Pax2, was identified along with a motif for Hnf1. Aside from roles in kidney, Hnf1 is also known to regulate genes in the pancreas and the liver [65]. Although a role for the repressor Cutl1 has not been clearly described in liver, its binding motif appeared to be conserved in liver relative to other tissues [186,187]. Notably, none of the novel motifs defined based on classification ability were significantly enriched based on the strict conservation metric. In particular, conservation did not support the novel motif which was the only motif identified in mES c1. In general, promoters with mES enriched activity were characterized by a dearth of over-represented motifs, known and novel, relative to adult tissues. Although our limited motif results in mES cells may reflect the bias of existing motif databases and the limitations of

our motif-analyses strategies, we posit that long-range or distal regulatory elements might play a more critical role in regulating the expression of enriched transcripts in ES cells.

Although in general there are close associations among Pol II binding, histone modifications, and transcript levels at most tissue specific promoters, we find enrichment of "active" epigenetic marks at a number of promoters with weak to undetectable Pol II occupancy. This trend is particularly apparent for roughly half of the promoters with enriched Pol II binding and gene expression in mES (mES c2). These promoters remain epigenetically marked by H3ac and H3K4me3 in adult tissues. Modifications associated with transcriptional activity, in particular H3K4me3, have been suggested to play additional roles as markers of recent transcription or poised activation at promoters, directly or indirectly inhibiting other forms of chromatin-mediated repression [188-192]. Subtle differences in the known function and identity of genes between the two mES classes reveal more known mouse embryonic stem cell markers within mES c1 (*Nanog, Pouf51, Dppa4, Nr0b1, Utf1, Tdgf1*). Promoters in mES c2 might be associated with a unique set of genes, such as the Gli zinc finger transcription factors, expressed at low levels, or in a small subset of cell types, within adult tissues [169]. The mES c2 category, relative to its complement among promoters with mES enriched activity, is distinguished by a two-fold higher overlap with CpG islands (45%). This sequence distinction might provide a clue to understanding this class and its regulation [191,192]. Further work is under way to more precisely characterize this phenomenon and its extent.

Our approach toward understanding tissue-specific gene expression integrates Pol II binding, chromatin modifications, and sequence features of promoters with measurements of relative transcript abundance. The genomic maps of Pol II binding and

chromatin modifications will be valuable resources that complement profiles of transcript levels and abundance for unraveling the layers of control governing gene expression patterns across cell types. Mapping of these features at different cell types at various developmental stages will likely provide further insight as to how cell-specific programs of expression are specified by sequence and epigenetic features across development.

4.6 Methods

Sample Preparation

R1 ES cells (a gift from Dr. Don Cleveland, Ludwig Institute for Cancer Research, San Diego) were maintained on top of feeder cells in cell culture dish with DMEM high glucose medium supplemented with 15% FBS, 0.1mM non-essential amino acid, 1mM sodium pyruvate, 1μM β-mercaptoethanol, 2mM L-glutamine, 50g/ml pen/strep and LIF. Cells were passed once on 0.1% gelatin without feeder cells before harvested. Cells were harvested and crosslinked with 1% formaldehyde for 20 minutes when they reached ~80% confluence on the plates. Mouse tissues were dissected from a 10-12 week old female BL6 mouse, chopped into small pieces (about $1mm^3$) with a razor blade in cold 1XPBS, and crosslinked with 1% formaldehyde for 30 minutes at room temperature. Cells were then sonicated as described in Z. Li and colleagues[64]

Chromatin Immunoprecipitation with Microarrays (ChIP-chip)

Chromatin immunoprecipitation was performed as previously described [64]. Briefly, 2 mg of sonicated chromatin (OD_{260}) was incubated with 10 μg of antibody (anti-RNA polymerase II, MMS-126R, Covance; anti-H3ac, 06-599, Upstate; anti-Me3H3K4, 07-473, Upstate) coupled to the IgG magnetic beads (Dynal Biotech). The magnetic

beads were washed eight times with RIPA buffer (50 mM Hepes at pH 8.0, 1 mM EDTA, 1% NP-40, 0.7% DOC, and 0.5 M LiCl, supplemented with Complete protease inhibitors from Roche Applied Science), and washed once with TE (10 mM Tris at pH 8.0, 1 mM EDTA). After washing, the bound DNA was eluted at 65°C in elution buffer (10 mM Tris at pH 8.0, 1 mM EDTA, and 1% SDS). The eluted DNA was incubated at 65°C overnight to reverse the cross-links. Following incubation, the immunoprecipitated DNA was treated sequentially with Proteinase K and RNase A, and was desalted using the QIAquick PCR purification kit (Qiagen). The purified DNA was blunt ended using T4 polymerase (New England Biolabs) and ligated to the linkers (oJW102, 5'-GCGGTGACCCGGGAGATCTGAATTC-3', and oJW103, 5'-GAATTCAGATC-3'). The ligated DNA was subjected to ligation-mediated PCR, labeled with Cy3 and Cy5 dCTP using a BioPrime DNA labeling kit (Invitrogen), and hybridized to the mouse genome tiling microarray.

The 37 genome-scan tiling array set containing 14.5 50-mer oligonucleotides, positioned at every 100 bp were designed and fabricated using the maskless array synthesis technology (MAS) by NimbleGen Systems. These arrays were designed to contain all the non-repetitive sequences throughout the mouse genome (NCBIv33, mm5).

Initial Identification of Pol II Binding Sites in Five Tissues

After scanning and image extraction, Cy5 (ChIP DNA) and Cy3 (input) signal values for each of the 37 genome tiling arrays were normalized by intensity-dependent Loess using the R package *limma* [85,193]. Median filtering (window size=3 probes) was used to smooth $\log_2$ (Cy5/Cy3) data across the tiled regions. For each array, ChIP-enriched probe clusters were defined as regions with a minimum of 4 probes separated by a maximum of 500 bp

with filtered $\log_2$R greater than 2.5 standard deviations from the mean log ratio, as used in our previous study of TAF1-binding in the human genome [43].

The application of the analysis above for each genome-scan tiling set corresponding to Pol II ChIP-chip for each tissue resulted in five sets (brain, heart, kidney, liver, embryonic stem cells) of putative Pol II binding regions in the mouse.

Condensed Array ChIP-chip

We designed a condensed array by combining the five sets of putative Pol II binding regions from the five Pol II genome-wide scans. Each binding region was extended by 2000 bp upstream and downstream and overlapping regions from the Pol II ChIP-chip of different tissues were merged to yield a set of 32,482 putative Pol II binding regions for condensed array design. NimbleGen Systems used the same probe designs from the genome-scan tiling set overlapping the 32,482 regions to synthesize the condensed scan array set containing 1.5 million probes in 4 arrays.

We performed 15 ChIP-chip experiments over the condensed array design for 3 factors (Pol II, H3ac, H3K4me3) across five mouse tissues. After scanning and image extraction, Cy5 (ChIP DNA) and Cy3 (input) signal values for each of the 4 condensed-scan tiling arrays (in each set) were normalized by applying either intensity-dependent Loess or median-scaling normalization with the correction based only on the intensities of 14,572 control probes (designated RANDOM_GC11_GC34). The R package *limma* was used to implement the normalization. Prior to comparison and clustering, array data were quantile normalized across tissues using the normalize.quantiles function in the R package *affy* [85,193].

Final Catalog of Pol II Binding Sites

To define a final catalog of Pol II binding sites we applied an improved version of the peakfinding algorithm which we previously used to define Taf1 binding in human IMR90 cells [43,72]. This algorithm predicts a binding site for a factor at the probe-level resolution. The *p*-value for significant peaks is based on the following test-statistic:

$$\hat{Y}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

Here *n* is the number of probes in the window forming a triangle centered at the predicted peak; Y_i is the log ratio for probe *i* within the window. The algorithm does not use a pre-specified window size but computes the statistic for all possible windows of a certain size range containing triangles centered at the predicted peak. We chose a *p*-value cutoff of $p<0.05$ to define significant peaks for Pol II binding in both the condensed scan and genomewide scan for each tissue. We designated a peak in the condensed scan as confirmed if the peak is predicted within 500bp of the peak identified in the genome-wide scan for each tissue. We define the coordinates of the confirmed peaks as the range defined by the matching condensed scan peak and genome scan peak.

As a second step in defining a catalog of Pol II binding sites, we pooled the confirmed peaks in each tissue and merged all the sites that are within 1000 bp of each other. This cutoff was based on the distribution of nearest neighbor distances between confirmed peaks. Sites were then merged across tissues if there was any base pair overlap. The Pol II binding site is then defined as the range of the confirmed peaks merged across tissues.

Expression Analysis

To complement the Pol II mapping strategy, we defined the set of genes with

transcripts relatively enriched in each tissue. We identified these genes by analyzing the genome-wide expression profiles of the each tissue using Affymetrix GeneChIP mouse 430 2.0, which represent over 39,000 mouse transcripts. Total RNA from each mouse tissue was extracted using Trizol® reagent (Invitrogen, Carlsbad, CA) and further purified using RNeasy Mini Kit (Qiagen, Valencia, CA) according to manufacturers' recommendations. The purified total RNA was submitted to UCSD Cancer Center Microarray Resource for GeneChip® RNA Expression Analysis using mouse 430 2.0, arrays. The resulting hybridization data was analyzed using Affymetrix GCOS v. 2.0 to determine the detection call as present (P), marginal (M), or absent (A) at significance level $p<0.05$.

We used annotation from the Affymetrix library file Mouse430_2.cdf to match probe sets to corresponding Entrez gene identifiers. Probe sets with identifier extension "x_at" were removed from the analysis. A total of 20,827 Entrez genes were mapped to the remaining probe sets. We performed quantile normalization on the probe set signals across tissues using the R package *affy*. To assign a signal for a gene in each tissue, we selected the maximum normalized expression signal of all probe sets matched to the gene if there are multiple probe sets for a gene. Tissue-specific measures of entropy and categorical tissue-specificity based on expression were computed as previously described[27].

Quantitative PCR

Quantitative real-time PCR was performed with 0.5 ng of H3ac, Me3H3K4 or RNA polymerase II ChIP DNA and total genomic DNA, as described previously. The quantitative real-time PCR of each sample was performed in triplicate using iCycler™

and SYBR green iQ™ SYBR green supermix reagent (Bio-Rad Laboratories). The threshold cycle (Ct) values were calculated automatically by the iCycle iQ™ Real-Time Dectection System Software (Bio-Rad Laboratories). Normalized ΔCt values for each sample were then calculated by subtracting the Ct value obtained for the unenriched DNA from the Ct value for the ChIP DNA ($\Delta Ct = Ct_{ChIP} - Ct_{total}$). The fold enrichment of ChIP DNA over the unenriched DNA was estimated using the formula $2^{-(\Delta Ct)}$.

Reporter assays

500 bp DNA fragments (250 bp upstream and downstream of RNA polymerase II peak) were cloned into the pGL3Basic plasmid (promega), in front of the promoterless firefly luciferase gene. 200ng of these plasmids were cotransfected with 2ng of pRL-CMV, a renilla luciferse reporter, into mouse hepatocyte AML12 (ATCC # CRL-2254) using lipofectamine2000 (Invitrogen). Transfected cells were harvested 48 hours after transfection, and Luciferase activity was measured using the Dual Luciferase Kit from promega according to vendor protocol. The ratio of firefly luciferase to renilla luciferse was used as the relative activity for each sample. 5 RNA polymerase II intragenic DNA fragments were used as negative controls.

Comparison of Pol II Binding Sites with Transcriptional Start Sites and Known Genes

Annotation Source	**Number of Transcripts**
UCSC Genome Browser MM5 (downloaded April 2006) [93]	
refGene*	19,363
knownGene	36,838
ensGene	31,035

all_mrna*	156,546

*Filtered to transcripts with >95% similarity to matching genomic loci.

We compared the location of 24,363 Pol II binding sites to annotated 5' ends from the UCSC Genome Browser by comparing the distance of the 5' end to the closest edge of a binding site. Each Pol II binding site was then matched to its closest transcript within 2.5kbp.

Transcript-matched Pol II binding sites were matched to mouse Entrez gene identifiers using the gene2accession table downloaded from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA, and the remainder by using the mouse gene annotation derived from http://symatlas.gnf.org.

Comparison of Pol II Binding Sites with RIKEN CAGE data

We used a similar strategy to compare binding sites with RIKEN CAGE data. The data file (rikenCageTc.txt), containing 594,136 CAGE tag clusters, was downloaded from the UCSC Genome Browser MM5 in July 2006.

Genomic distribution assignment

We assessed the genomic distribution of sites not matched to known TSS (n=7,387) by comparing their position relative to known gene and transcript annotation. Sites were assigned to a genomic distribution class given the following criteria and order of assignment: a) Exonic: If the site directly overlaps an exon, it is classified as exonic. Sites which are less than 1Kbp in size were extended to 1Kbp in total length (equal extension on each side), before assessing overlap. Exon coordinates were downloaded from the UCSC Genome Browser (MM5) based on the block information for mRNA and EST tracks (Table Browser). b) Intronic: If the site is within a transcript but not

overlapping exon annotation, it is classified as intronic. c) 3' Proximal: If the site is within 2.5Kbp downstream of the 3' end of a gene, it is classified as 3' Proximal. d) 5' Distal: If the site is more than 2.5 Kbp upstream but within 100Kbp upstream of a gene 5' end, it is classified as 5' Distal. e) 5' Distal: If the site is more than 2.5 Kbp downstream but within 100Kbp downstream of a gene 3' end, it is classified as 3' Distal. F) Gene Desert: The remainder of the sites that do not fall within gene or transcript boundaries and are not within 100Kbp of the boundary annotations are classified as falling within "Gene Deserts". The gene distal and gene desert classes were adapted from another publication [154].

Promoter Prediction Criteria

We also applied the peakfinding algorithm for the histone modification ChIP-chip data (H3ac, H3K4me3) in each tissue in order to define sites of histone modification enrichment at $p<0.05$. Given the observed association of H3K4me3 with the 5' end of genes from genome-scale studies in yeast, we classified CAGE and TSS-unmatched Pol II binding sites within 1000bp of H3K4me3 enrichment (in the same tissue) as promoters.

Comparison of Pol II Binding Sites with microRNA annotation

MicroRNA annotation was downloaded from miRBase (Release 8.1, May 2006) [158]. MM8 coordinates for 336 miRNAs were converted to MM5 coordinates in two steps using UCSC Genome Browser Utility liftOver and conversion tables (1) mm8toMm7.over.chain and (2) mm7toMm5.over.chain downloaded in May 2006 [93]. 334 miRNA coordinates were converted to mm5 and matched to Pol II binding sites. 85 miRNAs mapped within 2.5kb of 118 Pol II binding sites. 66 of 85 miRNAs were also

within 2.5 kb of annotated transcription start sites (TSS) while 19 miRNAs were outside 2.5kb of a known TSS.

Comparison of Pol II Binding Sites with other ChIP-CHIP and ChIP-PET Data

From Supplementary Table 2 (Table S2) of the article "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells," we obtained the MM5 coordinates of the binding sites for Oct4 and Nanog in E14 mouse ES cells [154]. We compared these coordinates with the location of our Pol II binding sites (not mapped to TSS, CAGE or promoter predictions) and matched those which are within 2.5 kb of the Pol II binding sites.

From Supplementary Table 9 (Table S9) of the article "Polycomb complexes repress developmental regulators in murine embryonic stem cells", we obtained the list of transcription factors bound by Polycomb complexes in mouse embryonic stem cells [164]. We compared the identifiers of the genes in the table with the genes in our tissue-specific sets.

Assessment of Tissue-Specificity

Each of the 24,363 sites was scored for tissue-specificity by calculating the Shannon entropy of the site's Pol II binding probability distribution across tissues, H_s. As described in 4.3.5, the relative Pol II binding probability in a tissue t for a given site s is calculated as:

$p_{t|s} = B_{t,s} \Big/ \sum_{1 \le t \le N} B_{t,s}$ where $B_{t,s}$ is the average ChIP-chip log ratio in the 1kb neighborhood centered at the midpoint of Pol II binding site s and N is the total number of tissues surveyed.

As described in results, H_s is computed for each site, and the categorical tissue-specificity $Q_{s|t}$ is calculated for each site and tissue combination to determine the bias of a site for a particular tissue[27].

The maximum entropy score is attained when the Pol II binding is uniform across tissues ($p_{t|s}$=1/5) for a given site. We used the arbitrary cutoffs of $H_s \leq 1$ and $\min(Q_{s|t}) \leq 1.32$ to define tissue-enriched sites. In the idealized case, $H_s \leq 1$ represents a probability distribution for a site *s* in which Pol II binding is predominantly enriched at ≤2 tissues (on average) with negligible relative occupancy at the rest of the tissues. Among all sites with a $\min(Q_{s|t}) \leq 1.32$, the maximum H_s is 1.01.

CpG Overlap of Pol II Binding Sites and Promoters

Any overlap of transcript-matched Pol II binding sites to CpG islands was determined by comparing the coordinates of CpG islands from the MM5 file *cpgIslandExt.txt* downloaded from the UCSC Genome Browser (MM5) [93].

CpG islands within tissue-specific gene promoters were identified based on the commonly used criteria requiring regions greater than 200bp in length, within [-200,+100]bp of the reference start, G or C content greater than 50%, and a ratio of observed over expected CG dinucleotide counts greater than 0.6 [194].

Assessment of Multiple Promoter Usage

We considered the set of genes with Entrez identifiers mapped to unique loci in the mouse genome and classified genes into two general classes based on the number of TSS-matched Pol II binding sites it contains (single promoter or multiple promoters). The fraction of genes with multiple promoters (28%) represents our conservative estimate

of multiple promoter usage given the limited panel of tissues surveyed and the coarser resolution of transcript 5' ends compared to RIKEN CAGE data.

To assess the relationship between the tissue-enrichment of multiple Pol II binding sites and relative tissue expression, we examined all genes (based on Affymetrix gene loci annotation) with ≥2 Pol II binding sites (n=3,843). We calculated the correlation coefficient (pearson R) between the relative tissue-binding vector (Pol II) and relative tissue expression vector for each gene. The relative tissue-binding vector of a gene contains the aggregate $B_{t,s}$ for each tissue across all the sites *s* matched to the gene *G*:

$$\left[\sum_{s \in G} B_{brain,s} , \sum_{s \in G} B_{heart,s} , \sum_{s \in G} B_{kidney,s} , \sum_{s \in G} B_{liver,s} , \sum_{s \in G} B_{mES,s} \right]$$

The relative tissue-expression vector for each gene *G* is simply the associated normalized log10 signal (expr) from the Affymetrix expression array experiment for each tissue:

$[expr_{brain,G}, expr_{heart,G}, expr_{kidney,G}, expr_{liver,G}, expr_{mES,G}]$

To assess the significance of the proportion of genes with strong positive correlation between relative tissue-binding and expression, we randomly permuted the labels of binding and expression vectors 1000 times and obtained an estimate of the expected proportion of genes across the range of correlation coefficients, $R \in [-1,1]$. We found that the expected distribution across correlation coefficients is relatively uniform and that the observed percentage of genes with R>0.6 (46%) between relative tissue binding enrichment and expression is >2-fold enriched above the expected (21%).

Assessment of Large Clusters of Tissue-Enriched Pol II Binding

From our catalog of Pol II binding sites, we selected 53 sites defined as tissue-enriched with length >5kb. 40 of these sites were matched to known genes with Entrez identifiers.

Evalutating Tissue-Specific Gene Promoters

We considered the set of genes which had TSS-matched Pol II binding sites enriched in only one tissue (max (H_s)≤1 and max($Q_{s|t}$)≤1.32 for only one tissue). We selected one reference start for each gene based on the Pol II binding site mapped closest to a TSS if there are multiple Pol II binding sites for a tissue. The reference start is defined differently for the following cases: 1) If the Pol II binding site contains the matching annotated transcript start (regardless of size), the annotated transcript start is used as the reference start. 2) If the Pol II binding site is less than or equal to1kb in span (and does not contain the matching annotated start), use the midpoint of the site as the reference start. 3) If the Pol II binding site is greater than 1 kb in span (and does not contain the matching annotated start), use the edge of the site closest to the matching annotated transcript start as the reference start.

Promoter Profile

For each set of genes with tissue-enriched Pol II binding, we extracted the ChIP-chip logR profiles for Pol II, H3ac, and H3K4me3 for all tissues over the [-2,+2]Kbp interval relative to the reference start. For each gene, we concatenated these "promoter profiles" for each factor, for all tissues. Each gene is represented by 15 concatenated vectors of 40 values, one for each 100bp bin in the [-2,+2]Kbp interval. For each set of tissue-enriched genes, we performed k-means clustering (pairwise complete-linkage) on

these concatenated tissue promoter profiles as implemented in the command line version of Cluster 3.0 downloaded from (http://bonsai.ims.utokyo.ac.jp/~mdehoon/software/cluster/software.htm) [100]. Unlike brain, heart, kidney, and liver which are predominantly enriched in Pol II and histone modifications in a tissue-specific manner, we observed that at k=2, the profiles for the mES set was roughly split into two major classes:

Class 1 (c1): A class with clearly mES-enriched Pol II binding, H3ac, and H3K4me3 enrichment relative to other tissues.

Class 2 (c2): A class with clearly mES-enriched Pol II binding, H3ac, and H3K4me3 enrichment but also detectable but weaker enrichment of H3ac, H3K4me3 (and barely detectable Pol II in average profiles) in other tissues. The TreeView application was used for visualization [101].

Expression Profile

We profiled the matching expression data based on our in-house Affymetrix experiments. Signals were logged ($\log_{10}$) and normalized across tissues for each gene.

We also downloaded the matching gene expression profiles across a panel of mouse tissues and cell types from SymAtlas (http://symatlas.gnf.org). We selected the matching data from the file *gnf1m-gcrma.txt* and performed hierarchical clustering (pairwise complete-linkage) of array (tissue) experiments using Cluster 3.0 on the gene mean-centered and normalized ($\log_{10}$) data before aligning to the tissue-set profiles to compare our results. The TreeView application was used for visualization [101].

Expression Correlation Score (ECS)

The metric to score the correlation of expression data with the tissue-enriched sets defined based on Pol II binding is a variant of the gene-set enrichment analysis as used by Xie, et al [24].

Given: Set of genes **S** independently defined as tissue-enriched by Pol II binding in a tissue **t**.

Goal: Evaluate the enrichment of the gene set **S** near the top of the ranked list **L** of genes ordered by tissue-specificity in **t** based on expression ($Q_{G|t}$). The categorical tissue-specificity score based on expression ($Q_{G|t}$) was calculated as described in the original application [27]. The ranked list **L** of genes includes all genes surveyed in the Affymetrix GeneChIP mouse 430 2.0.

Metric: Calculate the sum of the ranks of **S** in **L** (rank-sum statistic, R_S). Evaluate the non-randomness of ranks of **S** in **L** by comparing against the rank sums of random subsets (1000 random subsets) of **L** of the same size. Define the Expression Correlation Score (**ECS**) as:

$$ECS = \frac{\mu - R_S}{\sigma}$$

The ECS score is analogous to a z-score with an associated *p*-value. High SCS scores imply that the majority of items in **S** are near the top of the list **L**.

Tissue-specific sets defined based on Pol II binding alone are highly-enriched for genes which are highly tissue-specific based on expression data. Calculating the ECS

score for set S(t1) in ranked list L(t2) (where t1 and t2 are different tissues) shows significant anti-correlation (negative ECS score).

Motif Discovery

Data

Tissue-specific Pol II binding sites uniquely matched to known genes and within 2.5 kb of an annotated transcript start for the matching gene were used as anchors to define 1200bp proximal promoter regions spanning 1000bp upstream and 200bp downstream of a reference start. Reference starts are defined in the same way as for promoter profiles.

Random mouse promoters were selected from the Cold Spring Harbor Laboratory Mammalian Promoter Database (CSHLmpd). The promoter region was defined to be [-1000, +200] bp from the reference transcription start site [142].

Balanced Misclassification Metric

We identified motifs for each set of tissue-specific gene promoters by examining the relative over-representation of known vertebrate transcription factor binding site (TFBS) matrices based on Transfac [178] and JASPAR [179] (673) in each set compared to two types of background sets: 1) a random set of mammalian promoters or 2) the relative complement of the set in the set of all tissue-specific gene promoters. The mES c2 set was excluded from the relative complement sets of tissue-specific promoters because of its non-specific pattern of histone modification enrichment. A previously described enumerative strategy, DME, was also used to determine the highest ranked de novo discriminative motifs of different widths (w=6,8,10,12,14) in each tissue-specific set compared to each of the two types of background sets [25,144].

For both known and de novo motifs, a motif's ability to classify the foreground sequences from background sequences is measured by the balanced misclassification error rate. This error rate is defined as:

$$ErrorRate = 1 - \left[\left(Sensitivity + Specificity \right) / 2 \right]$$

Sensitivity is defined as the proportion of promoters in the foreground set containing the motif and specificity is defined as the proportion of promoters in the background set without the motif. The threshold for motif matching is optimized for each matrix to minimize the error rate.

The significance of the balanced misclassification error rate for a motif (*p*-value) is determined by estimating the expected distribution of the error rates for a given comparison. This is achieved by permuting the labels of foreground and background sequences 1000 times and calculating the misclassification error rate for the best discriminative motif for each permutation. The *p*-value for the top-ranked discriminative motif for the true foreground and background sets then represents the probability of obtaining a misclassification error rate less than the one observed based on the expected distribution of error-rates. This *p*-value is a conservative measure of significance when considering the error-rates of lower-ranked (rank >1) motifs for the true foreground and background comparison. For each comparison, we filtered the significant results to include motifs with *p*-value<0.05 and specificity>2/3.

Using tools from the CREAD (http://rulai.cshl.edu/cread/index.shtml) sequence analysis package (kmercomp, featuretab, and feateval), the ability of all possible short nucleotide sequences (k=1,2,3) to discriminate foreground and background sequences

were also compared with the error-rates of the motifs of greater widths (w>=6). By this strategy, we determined that in the comparison between the kidney promoter set versus random promoters, the "CAG" sequence is a better discriminator than the most significant motifs V$MYOD (M00184) and V$E12_Q6 (M00693). Similarly, in the comparison between the mES c2 set versus other tissue-specific promoters (brain, heart, kidney, liver), the "CG" dinucleotide is a better discriminator than the other significant CG-rich motifs from TRANSFAC and JASPAR (not reported). This reflects the differences in CG-composition between the mES c2 set and other tissues as reported in the analysis of CpG overlap.

Relative Tissue-Enrichment of Conserved Occurrences

Data

The 1200bp proximal promoter regions (MM5) were mapped to orthologous human promoters (HG17) in two steps by using UCSC utility liftOver and two conversion files: (1) mm5ToMm7.over.chain with the default minMatch cutoff ≥0.95 and (2) mm7ToHg17.over.chain with the minMatch cutoff ≥0.50. The following table summarizes the number of tissue-specific gene promoters mapped in human relative to mouse by this strategy:

Species	Brain	Heart	Kidney	Liver	mES c1	mES c2
Human	202	62	147	129	106	118
Mouse	219	70	174	159	157	158

Method

Given the set of known vertebrate TFBS matrices from TRANSFAC and JASPAR (678), the best occurrence of each motif was mapped at every promoter, for every tissue-specific set for both mouse and human using the CREAD (http://rulai.cshl.edu/cread/index.shtml) utility *storm*. Promoter occurrences for all motifs were filtered to those scoring above a functional depth threshold of 0.85:

$$FunctionalDepth = \frac{(Score - MinimumPossibleScore)}{(MaximumPossibleScore - MinimumPossibleScore)}$$

For every motif, we counted the number of promoters in which the best occurrence of the motif overlapped in the orthologous mouse and human promoters (aligned). We defined the total number of orthologous promoter pairs as *P*, the total number of orthologous promoter pairs with conserved occurrences of a motif *m* as *C*, the number of orthologous promoter pairs specific to the tissue as *T*, and the number of orthologous promoter pairs in *T* with conserved occurrences of the motif as *k*. We then scored the tissue-enrichment of the conserved occurrences for each motif (*m*) and for each tissue (*t*) by using the hypergeometric distribution [195].

$$p_{m,t} = 1 - \sum_{i=0}^{k} \frac{\binom{T}{i}\binom{P-T}{C-i}}{\binom{P}{C}}$$

p-values obtained from each of the 4,038 tests (673 motifs, 6 tissue sets) were classified as significant based on a *p*-value cutoff of $p<1/4038$ to account for multiple testing.

GO Analysis

We used the GO analysis tools at DAVID Bioinformatics Resources (http://david.abcc.ncifcrf.gov/) to determine the most enriched functional categories for the genes (based on Entrez identifiers) in each tissue set [95]. We focused on the GO-biological processes (BP) at level 3 (among levels 1-5) because of the intermediate level of descriptive information and category size. We highlight the top functional categories below the EASE score/DAVID *p*-value (a modified Fisher-exact *p*-value) of $p<0.05$ [95] based on the default genome-wide scope. There is no explicit multiple testing correction, but the ranking based on the score should reflect the most relevant categories for each tissue.

Acknowledgments

Chapter 4, in full is a manuscript prepared for submission as: Barrera, Leah O.; Li, Zirong; Smith, Andrew D; Zhang, Michael Q; Green, Roland; Ren, Bing. Genome-wide promoter profiling of mammalian tissues. I was a primary co-author and researcher of this work. I performed the computational portion and analysis of the research and wrote the paper. The other primary co-author performed all the experimental assays and validation. Other co-authors of the paper supervised and directed the work, edited the manuscript, or contributed analytical tools (DME *de novo* motif finder) and experimental materials (NimbleGen Arrays).

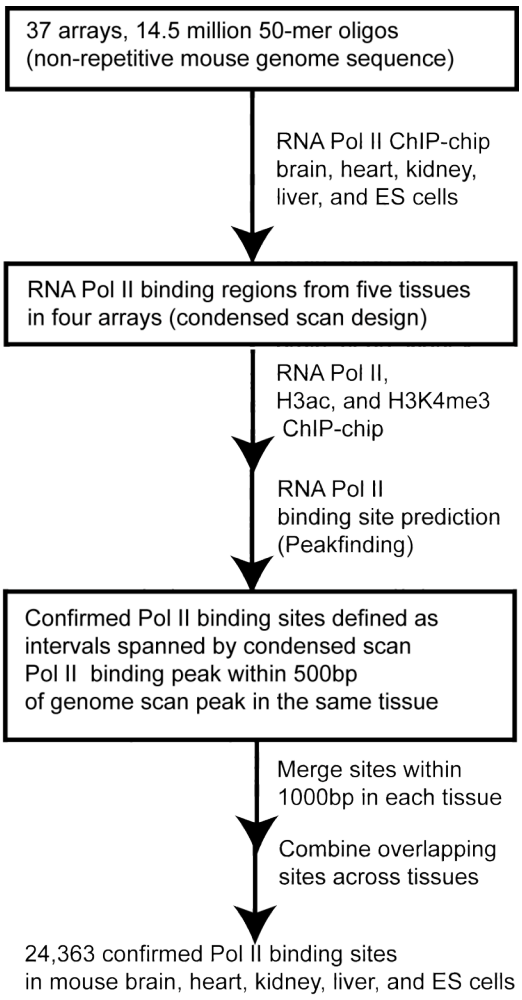


Figure 4-1. Outline of promoter mapping strategy.

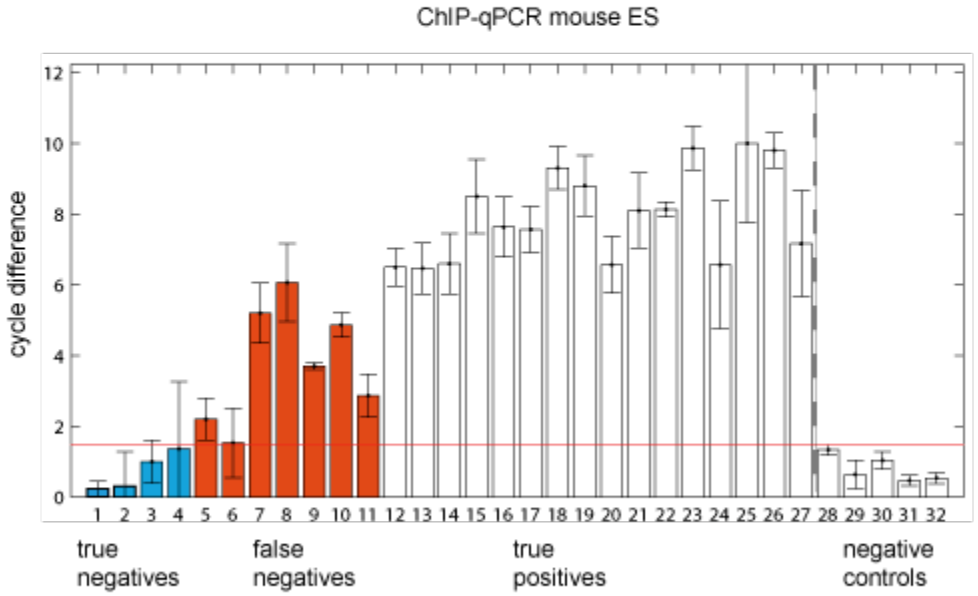


Figure 4-2. Summary of ChIP quantitative PCR analysis of 27 random RefSeq promoters in mES.

The y-axis denotes the cycle difference for ChIP enrichment relative to input at each site. Values for each bar represent the mean of 3 replicates with error bars showing the standard deviation. The first four promoters (L-R, on the x-axis) were determined to be relatively un-enriched in Pol II binding in mES by comparison with negative controls (last five promoters) selected from intergenic regions. Threshold for Pol II enrichment was defined as the mean of the negative controls plus three times the mean of the standard deviations for each negative control experiment.

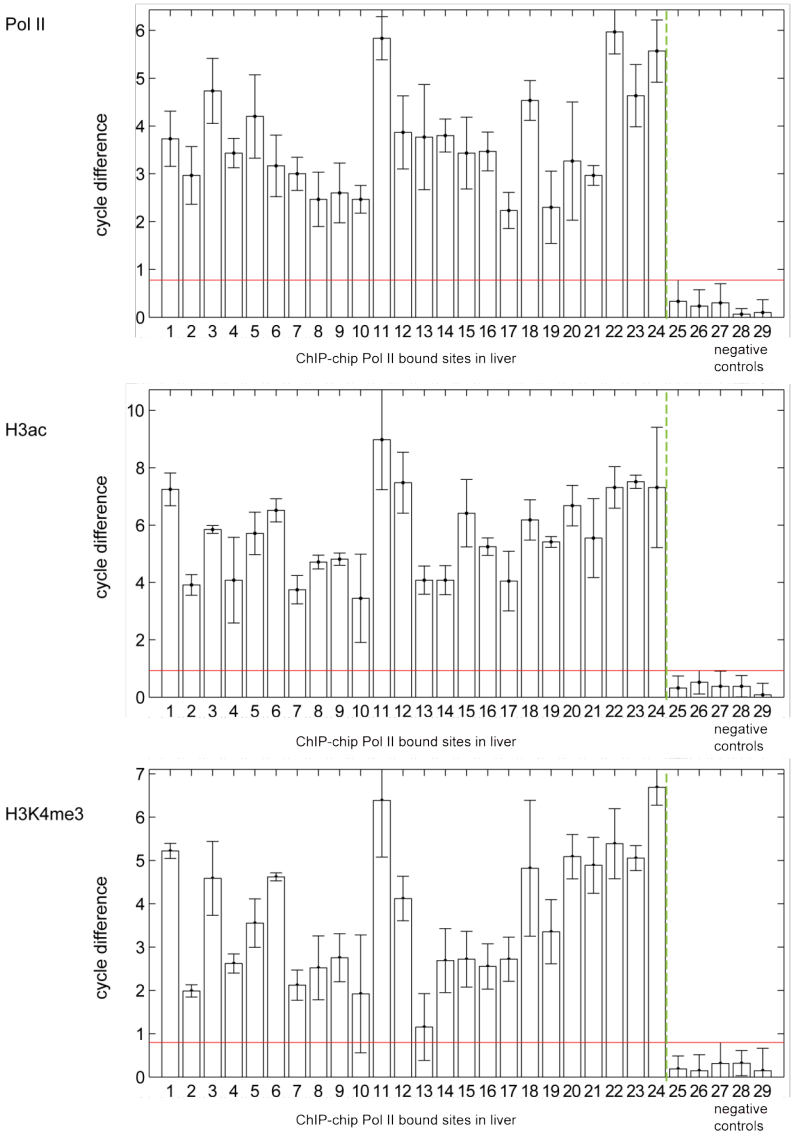


Figure 4-3. Summary of ChIP quantitative PCR analysis of Pol II, H3Ac, and H3K4me3 of 24 Pol II bound sites in liver.

The y-axis denotes the cycle difference for ChIP enrichment relative to input at each site. Values for each bar represent the mean of 3 replicates with error bars showing the standard deviation. The 24 sites tested in liver (L-R, on the x-axis) were compared with negative controls (25-29) selected from intergenic regions. Red horizontal line indicates maximum mean enrichment plus standard deviation for a negative control site.

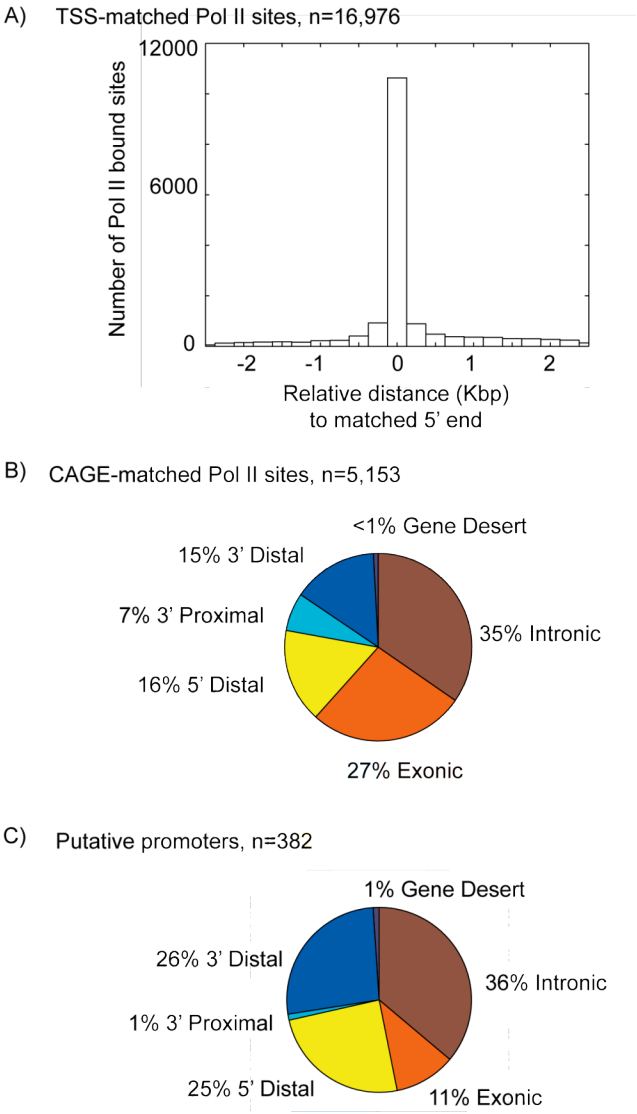


Figure 4-4. Genomic distribution of Pol II binding sites.

(A) Distance of matched Pol II binding sites relative to known 5' ends based on knownGene, refGene, ensGene, and all_mrna annotation downloaded from the UCSC Genome Browser (MM5). Bin size=250bp. (B) Genomic distribution of Pol II sites matched to CAGE annotation (and not matched near known transcript 5' ends). Genomic distribution criteria is as follows. 3' Proximal: within 2.5 Kbp downstream of 3' end. 5' Distal: 2.5Kbp to 100Kbp upstream of 5' end. 3' Distal: 2.5Kbp to 100Kbp downstream of 3' end. Exonic: overlapping exons (Pol II sites <1Kbp long extended to 1Kbp for overlap). Intronic: within transcript boundaries not near exons. Gene Desert: greater than 100kb from a 5' or 3' end. (C) Genomic distribution of putative promoters based on the criteria in B.

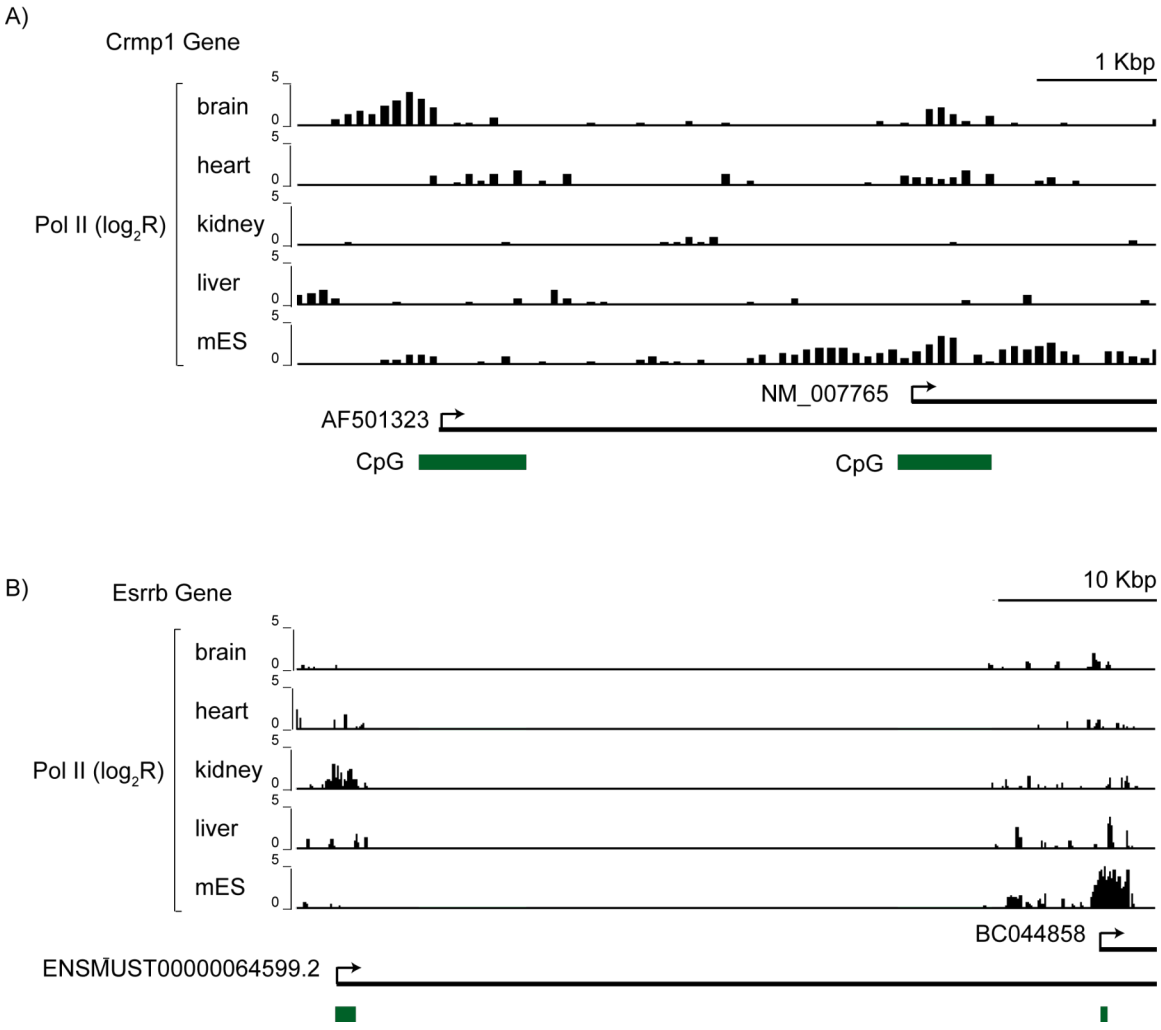


Figure 4-5. Examples of Pol II binding at promoters across tissues.

(A) Each bar represents the Pol II ChIP-chip $\log_2$ratio measured for each 50bp probe spaced by 50bp intervals over the region spanning the promoters for the Crmp1 gene. Differences in relative occupancy of Pol II across tissues suggest preferred usage of the AF501323 promoter in brain and conversely preferred usage of NM_007765 promoter in mES cells. Both promoters are shown to overlap CpG islands. (B) Pol II enrichment across tissues for the Esrrb gene. Alternative promoters are spread over a larger region. The upstream promoter (ENSMUST0000064599.2) has greatest Pol II ChIP-chip $\log_2$ratio enrichment in kidney while the downstream promoter (BC044858) has greatest Pol II ChIP-chip $\log_2$ratio enrichment in mES.

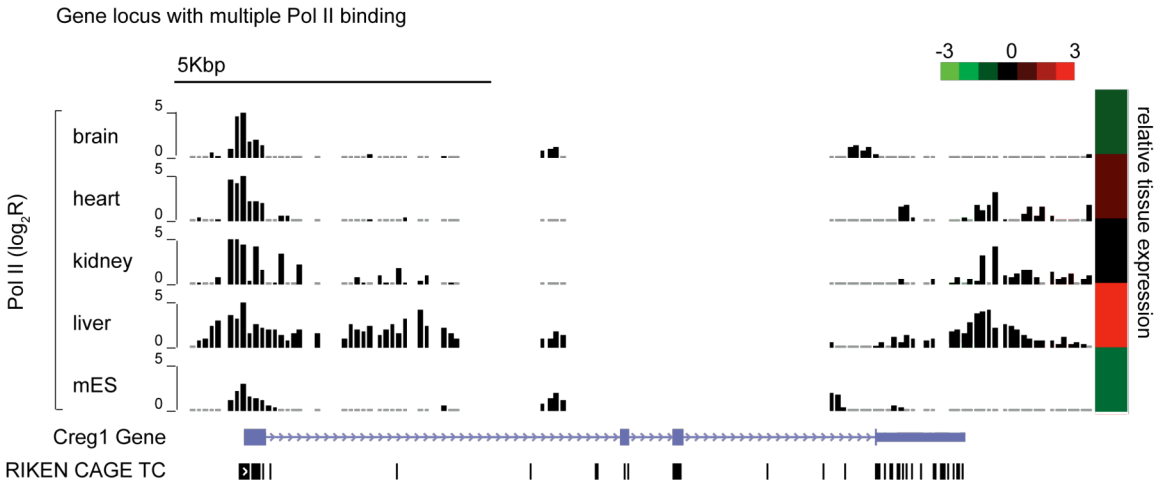


Figure 4-6. Gene locus with multiple Pol II binding sites across tissues and its relative expression.

The horizontal panels for each tissue represent Pol II ChIP-chip $\log_2$ratio enrichment across the gene loci. Each black bar represents the Pol II ChIP-chip $\log_2$ratio for a single 50bp probe. The vertical red-green bar at the far right represents the matching relative tissue-expression in the corresponding tissue based on normalized log10 signals from Affymetrix expression profiling.
Below the tissue panels, is the gene structure and orientation of the Creg1gene. At the bottom, we also match the RIKEN CAGE tag clusters (TCs) within the genomic window. Top-right scale shows relative expression enrichment associated with red, and the converse in green.

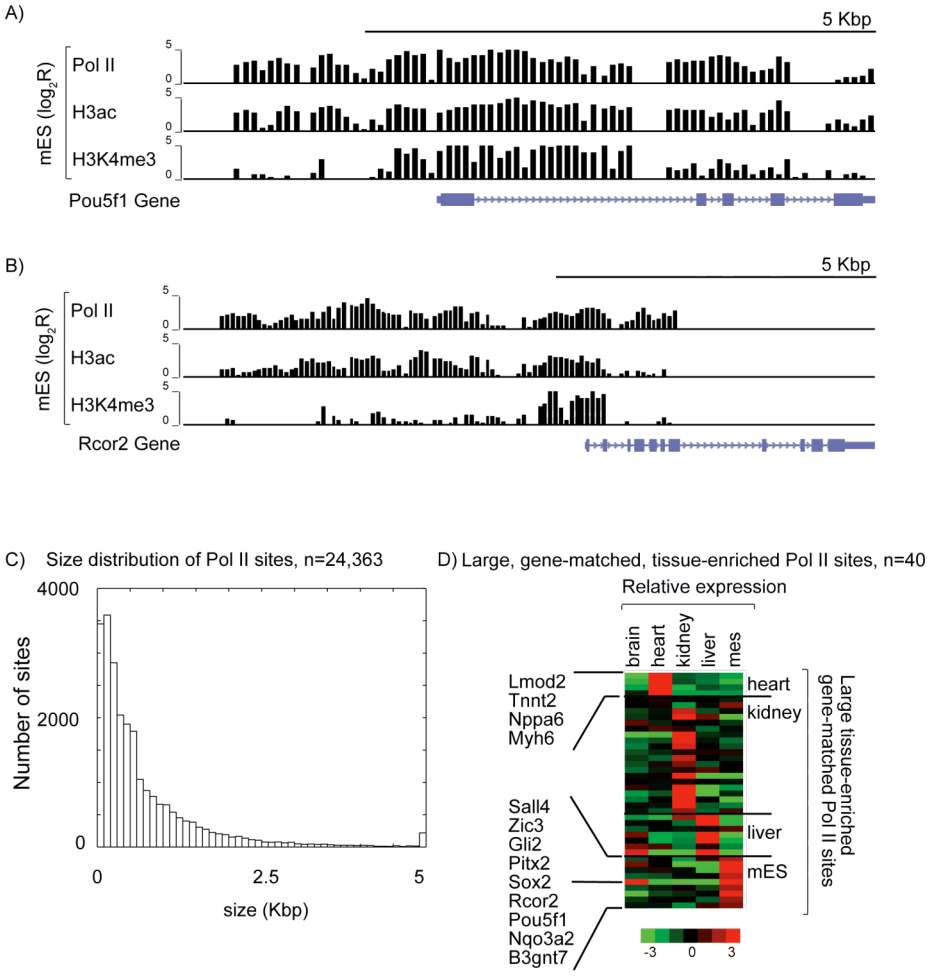


Figure 4-7. Unusually large regions of Pol II binding.

(A) Example of extended Pol II binding (>5Kbp) over the Pouf51 gene (chr17:34013130-34020755). We highlight the ChIP-chip $\log_2$ratio enrichment for Pol II, H3ac, and H3K4me3 in mES cells within each horizontal panels over the genomic window. Each bar represents the ChIP-chip $\log_2$ratio for the corresponding 50bp probe. The bottom panel shows the gene structure of Pou5f1 and its relative position and orientation within the window. (B) Example of extended binding over the Rcor2 gene (chr19:6976315-6984904). (C) Histogram of the size distribution of all 24,363 Pol II binding sites (bin size = 100bp). (D) Large Pol II binding sites overlapping known genes, n=40. Red-green heatmap of the relative expression (normalized log10 expression signal) across tissues for each matched gene. Genes are grouped (by row) according to the tissue in which Pol II binding is enriched. Red-green scale indicates relative expression enrichment associated with red, and the converse in green.

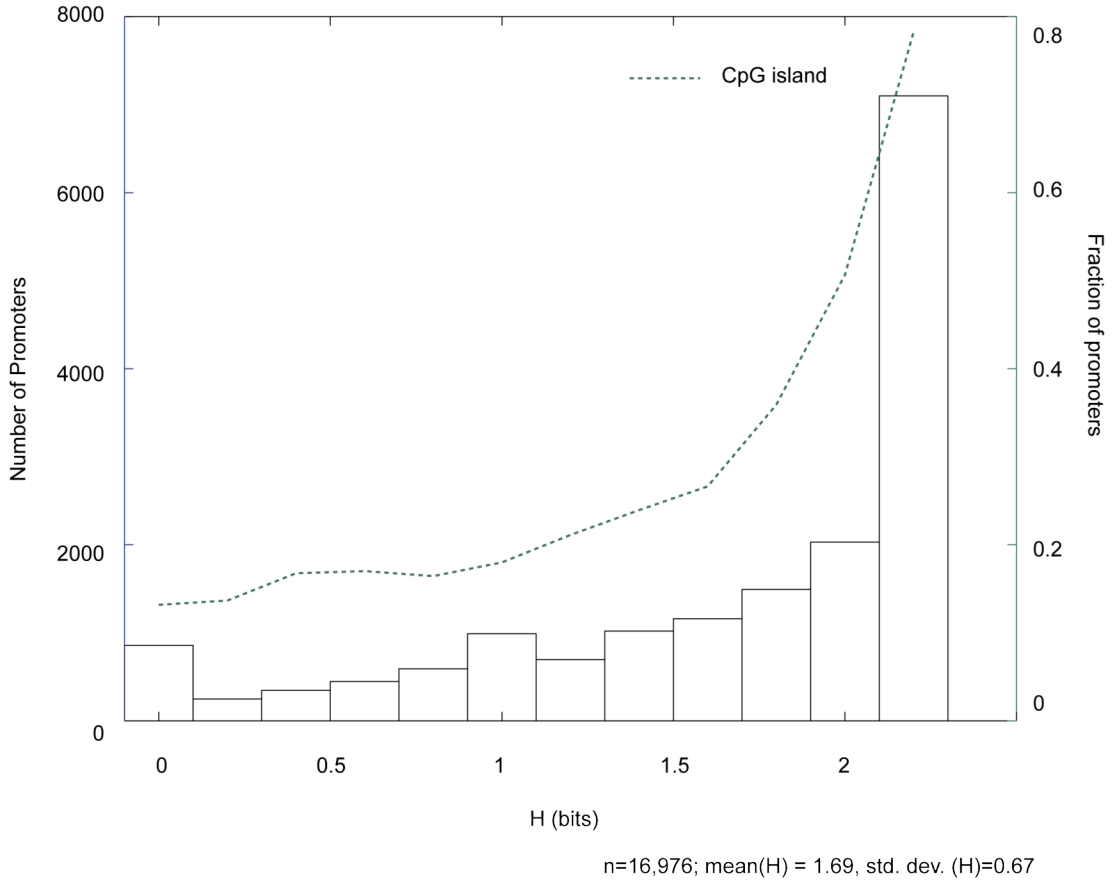


Figure 4-8. Tissue-specificity of known promoters based on Pol II binding and overlap with CpG Islands.

Bars indicate distribution of promoter counts (Y-axis, left) across the different bins (bin size = 0.2 bits) spanning the range of tissue-specificity measured by Shannon entropy (H) $H \in [0, \log_2(N)]$. Low values of H indicate tissue-specific expression and the maximal value denotes uniform expression across tissues surveyed. Dashed line indicates the fraction of promoters within each bin overlapping CpG Islands (Y-axis, right).

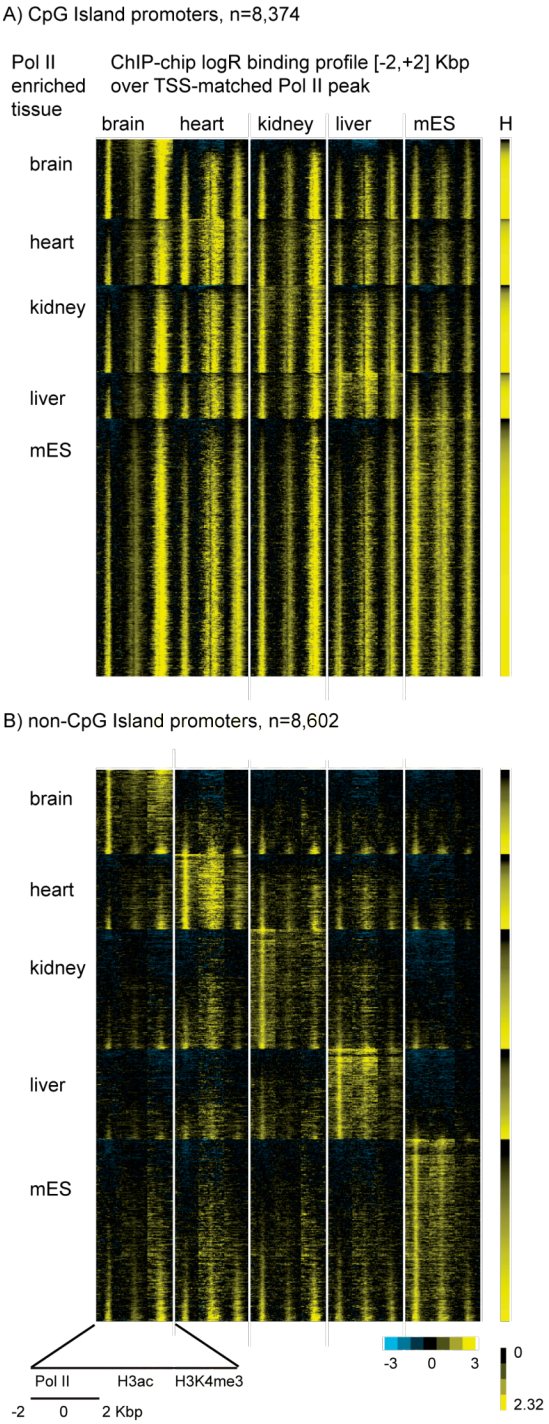


Figure 4-9. Promoter profiles of Pol II binding, H3ac, and H3K4me3.

Profiles grouped according to the tissue with the highest relative Pol II binding, and ordered within each tissue according to the Pol II entropy score or H (right bar) for all transcript-matched promoters (n=16,976). These promoters are partitioned by overlap with CpG islands: A) CpG promoters, n=8,374. B) Non-CpG promoters, n=8,602.

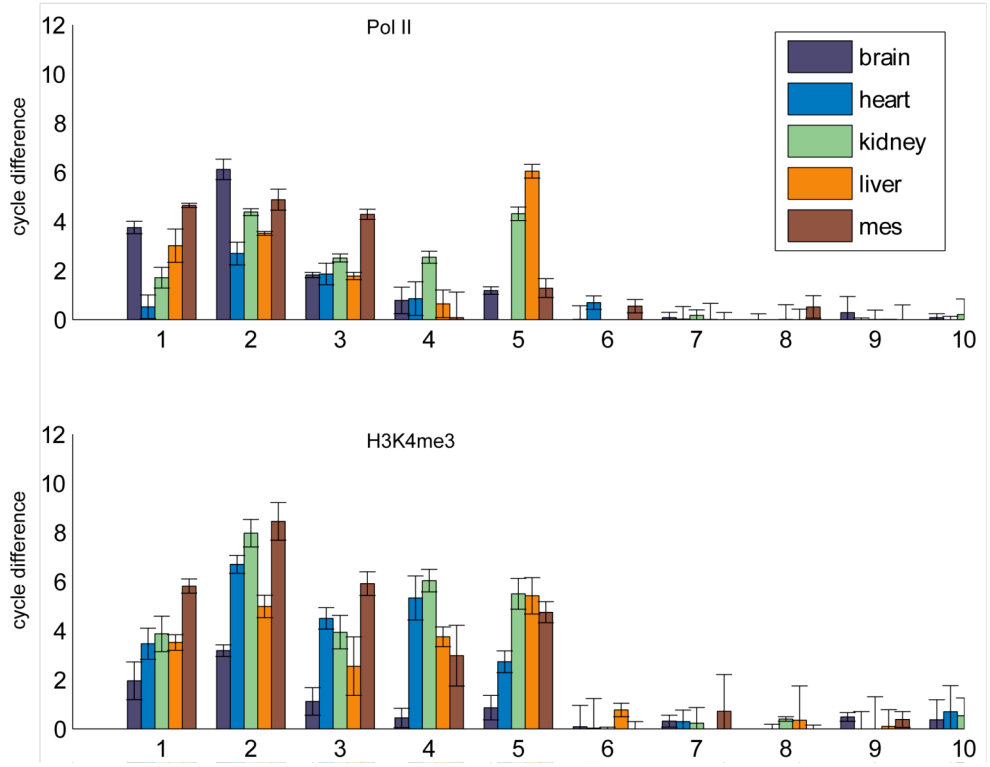


Figure 4-10. ChIP-qPCR validation of Pol II and H3K4me3 at 5 promoters.

Summary of ChIP quantitative PCR analysis of Pol II and H3K4me3 at 5 randomly selected promoters (1-5) with variable Pol II ChIP-chip binding across tissues and 5 negative control intergenic regions (6-10). The y-axis denotes the cycle difference for ChIP enrichment relative to input at each site. Values for each bar represent the mean of 3 replicates with error bars showing the standard deviation.

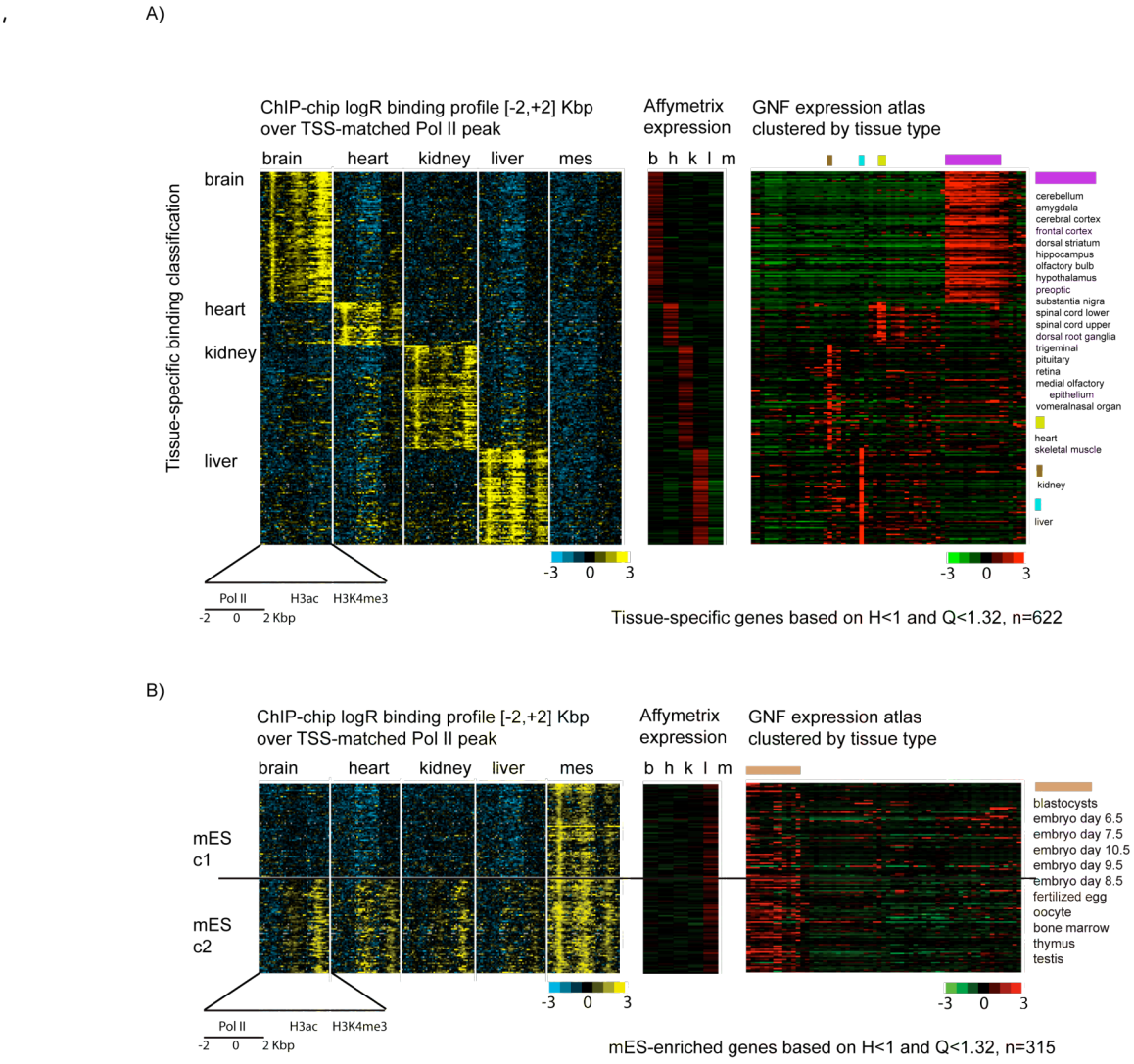


Figure 4-11. Tissue-specific gene promoter profiles and expression.

Pol II binding and histone modifications across tissues for tissue-specific promoters compared with relative transcript level based on expression profiling based on in-house Affymetrix expression of the tissues surveyed and relative to 61 tissues profiled in the GNF SymAtlas. Profile of Pol II binding, chromatin modifications at a tissue-specific promoter is concatenated across tissues and mapped to relative tissue expression of matching transcript based on Affymetrix expression profiling (in-house and GNF) along the same row. The rows are ordered and grouped according to the tissue-specific classification of the promoter based on Pol II binding.

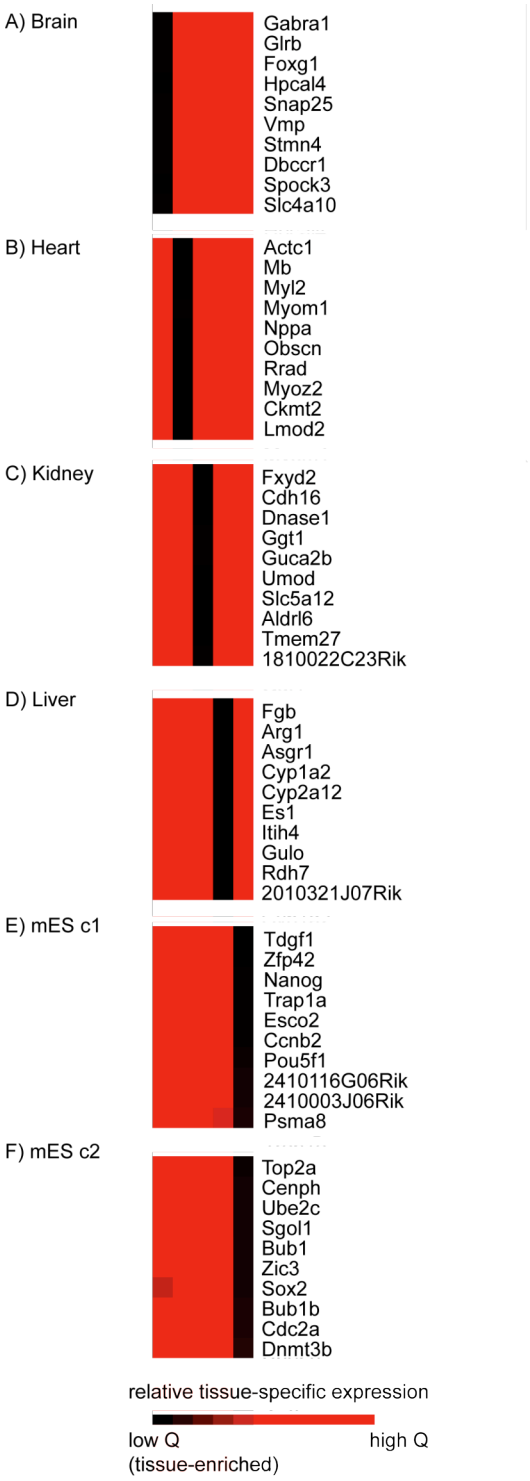


Figure 4-12. Tissue-specific genes based on promoter binding and relative transcript level.

We highlight the genes associated with tissue-specific promoters based on Pol II binding which rank highest in the list of genes ordered by categorical tissue-specific expression (top ten) for brain (A), heart (B), kidney (C), liver (D), mES c1 (E), mES c2 (F).

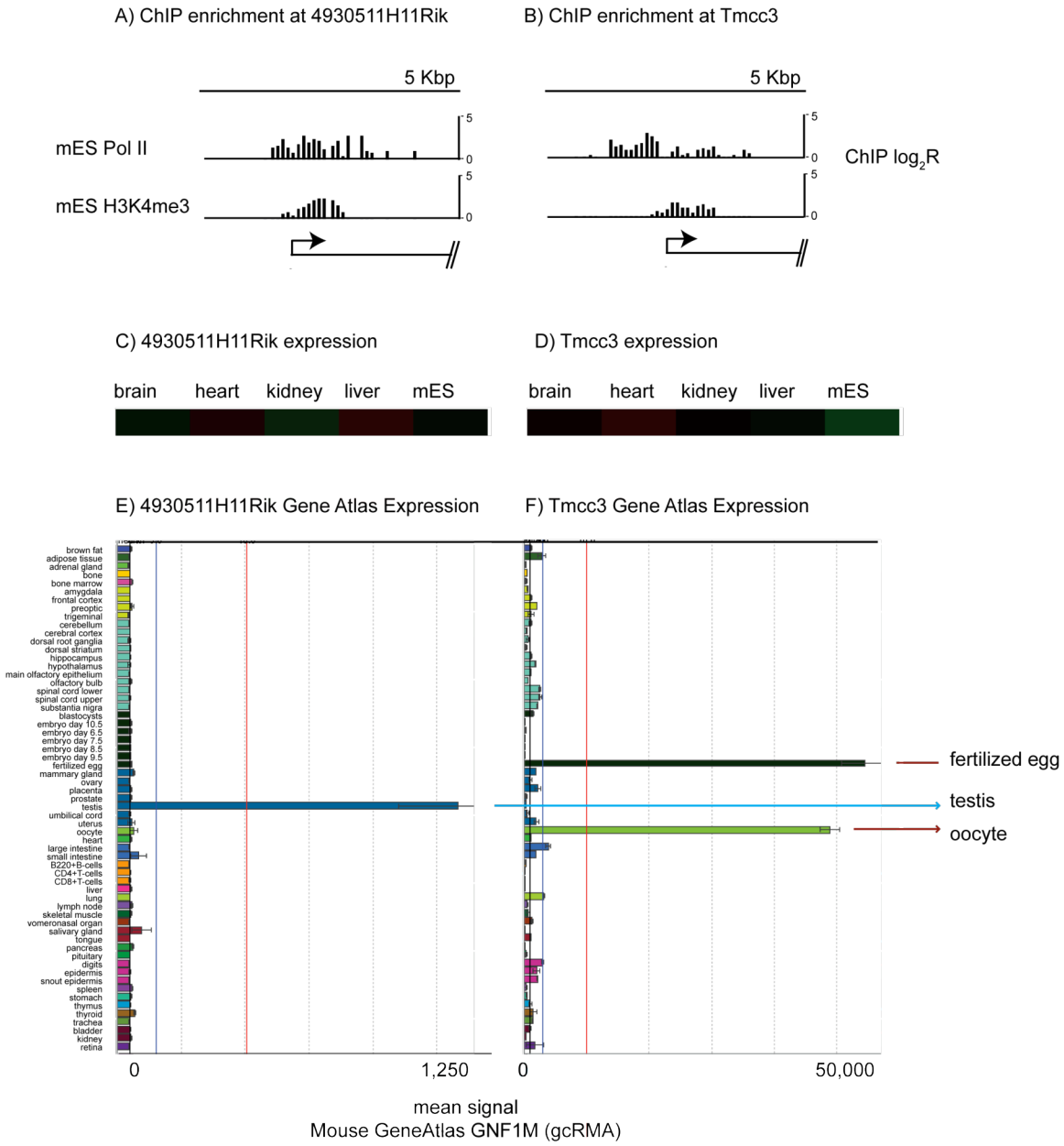


Figure 4-13. mES-enriched promoters with no transcript level correlation.

(A) Promoter profile for Pol II and H3K4me3 ChIP-chip $\log_2$ratio enrichment across tissues over the *4930511H11Rik* promoter. 5' end position (arrow) and relative gene orientation indicated by transcript schematic at the bottom. Each vertical bar represents the ChIP-chip $\log_2$ratio for the corresponding 50bp probe. (B) Promoter profile for the *Tmcc3* gene. (C) Relative expression of *4930511H11Rik* across the tissues surveyed based on normalized log10 signals from Affymetrix expression profiling. Expression enrichment from low to high is represented by color gradient from green to black to red. (D) Relative expression for the *Tmcc3* gene. (E) *4930511H11Rik* expression across a panel of cell types in the GNF expression atlas (copyright GNF) show enriched expression in testis. Each horizontal bar is a representative signal for each of the tissue type surveyed. (F) *Tmcc3* gene expression across a panel of cell types in the GNF expression atlas (copyright GNF) show enriched expression in fertilized egg and oocyte.

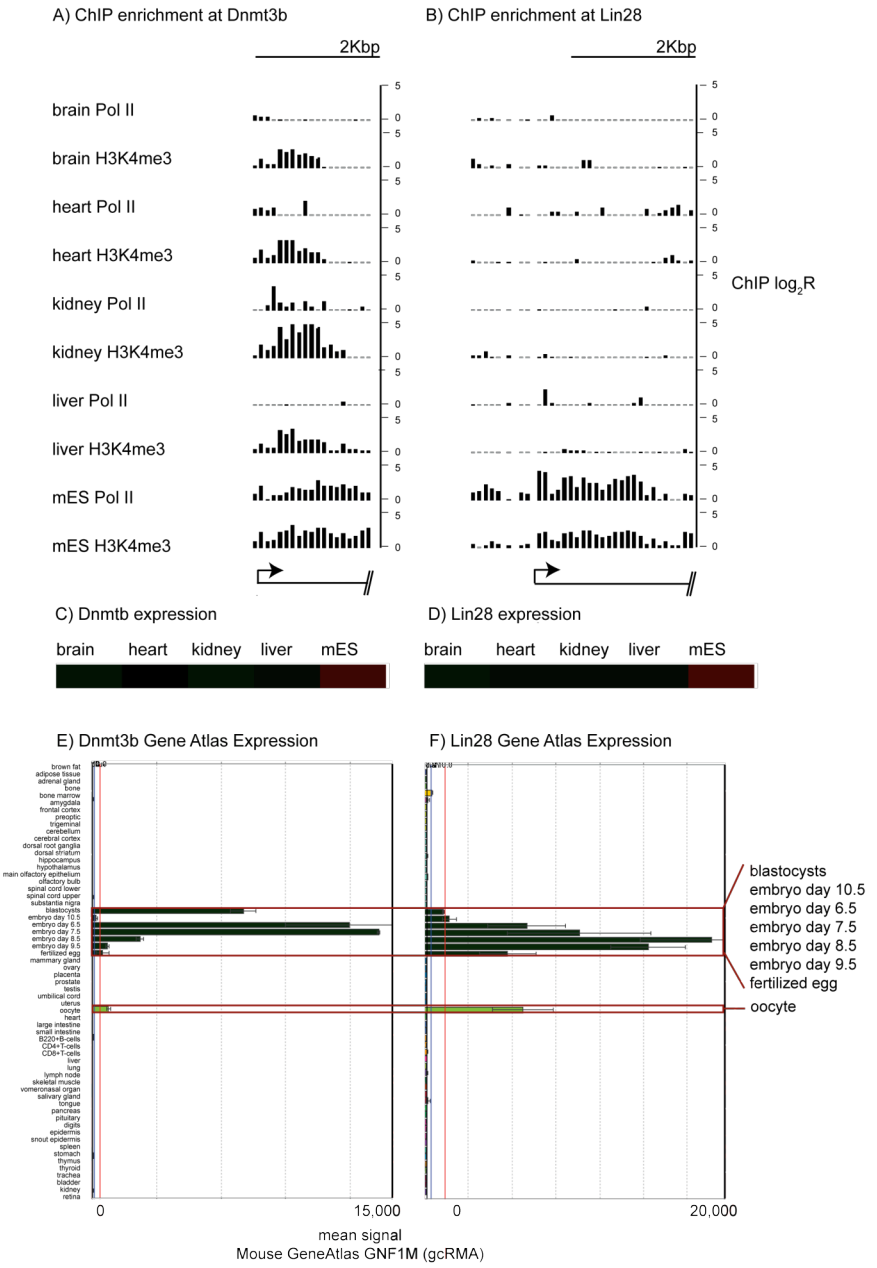


Figure 4-14. Transcript level and promoter profiles for mES c1 and mES c2.

(A) Promoter profile for Pol II and H3K4me3 ChIP-chip $\log_2$ratio enrichment across tissues over the Lin28 promoter. 5' end position (arrow) and relative gene orientation indicated by transcript schematic at the bottom. Each vertical bar represents the ChIP-chip $\log_2$ratio for the corresponding 50bp probe. (B) Promoter profile for Dnmt3b. (C) Relative expression of Lin28 across the tissues surveyed based on normalized log10 signals from Affymetrix expression profiling. Expression enrichment from low to high is represented by color gradient from green to black to red. (D) Relative expression for Dnmt3b. (E) Lin28 expression across a panel of cell types in the GNF expression atlas (copyright GNF) show enriched expression in embryo-related genes, fertilized egg, and oocyte. Each horizontal bar is a representative signal for each of the tissue type surveyed. (F) Dnmt3b expression across a panel of cell types in the GNF expression atlas (copyright GNF) show enriched expression in embryo-related genes, fertilized egg, and oocyte.

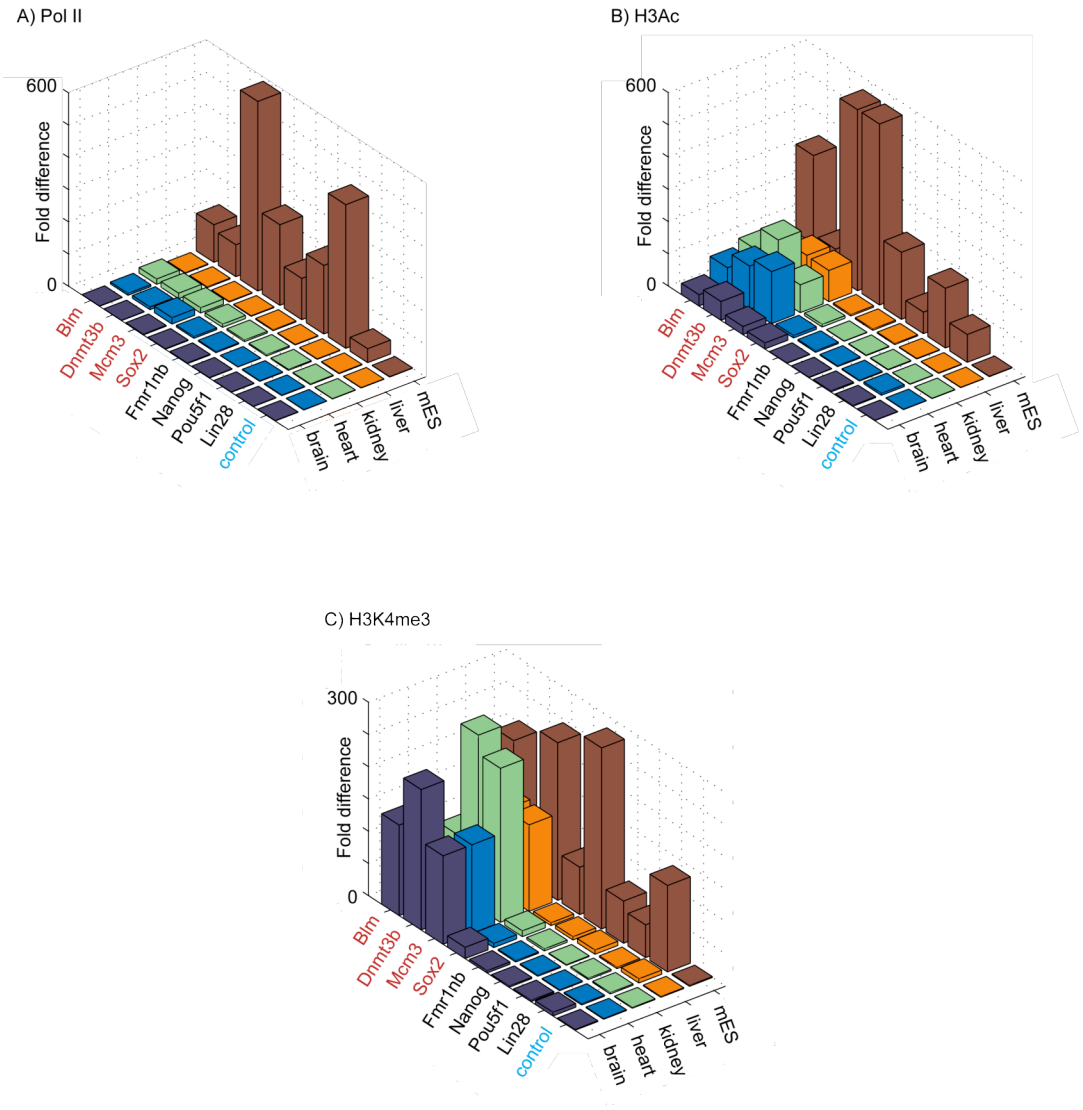


Figure 4-15. ChIP-qPCR validation of mES c1 and c2 classification.

(A) Average fold difference or Pol II ChIP DNA enrichment relative to input DNA (Z-axis) at the promoters for selected mES c2 genes (labeled in red), mES c1 genes (labeled in black), and an intergenic control (Y-axis) across the different tissues (X-axis). (B) Similar graph for H3ac. (C) Similar graph for H3K4me3.

Table 4-1. Summary of Pol II binding across tissues.

Pol II binding sites denote the number of sites associated with each tissue after merging the sites across tissues to define a total of 24,363 binding sites across tissues. Percent near TSS or CAGE is defined as being within 2.5 Kbp of the 5' end of the transcript or of the boundaries of the CAGE cluster.

	Pol II binding sites	Percent near TSS	Percent near TSS or CAGE
brain	8,173	86	96
heart	6,382	86	97
kidney	12,719	81	96
liver	9,127	78	94
mES	12,273	76	92
TOTAL	24,363	70	91

Table 4-2. Summary of Oct4 and Nanog co-localization.

Oct4 and Nanog co-localization with tissue-specific Pol II bound sites in mES versus adult tissues not mapped to promoters based on known transcript 5' end, CAGE, or H3K4me3 localization. Overlap between Pol II and Oct4 or Nanog binding is defined as within 2.5Kbp of their boundaries.

	MES	Other Tissues
Total unmatched tissue-enriched Pol2 sites	414	405
Co-localized Nanog (ChIP-PET, n=3006)	95	1
Co-localized Oct4 (ChIP-PET, n=1083)	32	0
Co-localized Nanog and Oct4	26	0
Total Co-localized Nanog/Oct4	101 (24%)	1 (0.2%)

-

Table 4-3. MicroRNAs matched to Pol II binding across tissues.

Genomic location is givenbased on the Pol II binding site. Highlighted in gray are microRNAs we found enriched in the same or related cell type as the cloning source.

miRNA ID	H	Most Enriched Tissue	mirBASE miRNA clone tissue sources	Genomic Location
mmu-mir-129-2*	0.82	brain	cerebellum	Outside Gene
mmu-mir-124a-3*	0.86	brain	brain,mES	Outside Gene
mmu-mir-9-3	1.15	brain	brain,mES	Intronic
mmu-mir-133a-2*	0.01	heart	heart	Intronic
mmu-mir-133a-1*	0.06	heart	heart	Intronic
mmu-mir-1-2*	0.49	heart	heart	Intronic
mmu-mir-681	0.67	heart	embryo	Intronic
mmu-mir-497	1.51	heart	embryo	Intronic
mmu-mir-145	1.69	heart	heart	Outside Gene
mmu-mir-143	1.73	heart	heart, spleen	Outside Gene
mmu-mir-23a	2.01	heart	heart	Intronic
mmu-mir-704	0.10	liver	embryo	Intronic
mmu-mir-122a*	0.32	liver	liver	Intronic
mmu-mir-190	0.74	liver	kidney	Intronic
mmu-mir-192	1.46	liver	liver	Intronic
mmu-mir-193	1.86	liver	kidney	Outside Gene
mmu-mir-469*	0.02	mES	testis	Outside Gene
mmu-mir-200c	1.19	mES	testis	Outside Gene
mmu-mir-202	1.60	mES	testis	Intronic

Table 4-4. Binding and expression correlation

Tissue-specific classification of gene promoters based on Pol II binding correlates with tissue-specific classification of genes based on expression using the metrics of Shannon entropy and categorical tissue-specificity. The expression correlation score (ECS) is a z-score measuring the enrichment of tissue-specific genes defined based on binding near the top of ranked lists for the same tissue ordered based on categorical tissue-specific expression (Affymetrix expression profiling).

tissue	genes	ECS	p-value
brain	219	19.63	<1e-16
heart	70	10.95	<1e-16
kidney	174	17.80	<1e-16
liver	159	18.71	<1e-16
mes1	157	10.11	<1e-16
mes2	158	14.08	<1e-16

Table 4-5. Summary of enriched Gene Ontology Biological Process (GO-BP).

Categories (Level 3) enriched for each tissue at p<0.05, ordered by increasing *p*-value.

Tissue	GO-BP	p-value
brain	cell-cell signaling	3.9E-10
	establishment of localization	1.7E-06
	transport	9.9E-06
	nervous system development	1.5E-05
	regulation of cyclase activity	8.5E-03
	regulation of lyase activity	8.5E-03
	neuron differentiation	2.5E-02
	cell motility	4.2E-02
	localization of cell	4.2E-02
heart	muscle contraction	9.9E-14
	circulation	3.6E-06
	muscle development	7.1E-04
	embryonic heart tube development	9.4E-04
	muscle cell differentiation	5.5E-03
	regulation of organismal physiological process	7.8E-03
	organ morphogenesis	1.1E-02
	embryonic development (sensu Metazoa)	1.4E-02
kidney	establishment of localization	5.3E-08
	transport	1.5E-07
	organ morphogenesis	3.2E-05
	ion homeostasis	7.2E-04
	cell homeostasis	2.9E-02
	skeletal development	3.3E-02
liver	transport	1.1E-06
	nitrogen compound metabolism	2.7E-06
	blood coagulation	3.1E-06
	regulation of body fluids	1.3E-05
	establishment of localization	1.3E-05
	response to pest, pathogen or parasite	1.7E-04
	immune response	2.0E-04
	response to other organism	2.4E-04
	catabolism	3.9E-04
	cell homeostasis	5.6E-04
	urea cycle intermediate metabolism	2.2E-03
	defense response	2.8E-03
	response to wounding	8.0E-03
	glucose homeostasis	1.1E-02
	cellular metabolism	1.7E-02
	regulation of organismal physiological process	2.4E-02

Tissue	GO-BP	p-value
mES c1	cell proliferation	5.2E-04
	regulation of cellular physiological process	1.7E-03
	anterior/posterior pattern formation	5.4E-03
	cell motility	6.9E-03
	localization of cell	6.9E-03
	embryonic development (sensu Metazoa)	8.9E-03
	negative regulation of physiological process	1.3E-02
	positive regulation of physiological process	2.1E-02
	embryonic pattern specification	2.5E-02
	cell growth	2.5E-02
	positive regulation of cellular process	2.6E-02
	morphogenesis of a branching structure	3.8E-02
	cell cycle	4.1E-02
	regulation of metabolism	4.8E-02
mES c2	cell cycle	5.6E-09
	cell division	3.9E-08
	chromosome segregation	4.9E-06
	primary metabolism	3.3E-05
	cellular metabolism	8.5E-05
	macromolecule metabolism	1.4E-03
	regulation of cellular physiological process	2.0E-03
	response to DNA damage stimulus	1.2E-02
	regulation of metabolism	1.2E-02
	embryonic development (sensu Metazoa)	1.2E-02
	DNA methylation	1.4E-02
	positive regulation of cellular process	4.1E-02

Table 4-6. Summary of known and novel motifs.

Identified in each tissue using a relative conservation metric and a balanced misclassification metric. Significant motifs identified using the relative conservation metric are based on a *p*-value threshold which takes into account the number of motifs and tissues tested (*p*-value cutoff <1/(motifs × tissues)). Error-rate *p*-values do not require multiple testing adjustment and are filtered at $p<0.05$.

Tissue	Factor(s)	Motif(s)	Selected reference(s)
Brain	Arnt-Ahr	MA0006	(Aitola and Pelto-Huikko 2003; Swanson et al. 1995)
	ATF	M00691|V$ATF1_Q6, M00017\|V$ATF_01, M00179|V$CREBP1_Q2	(Herdegen and Leah 1998)
	CREB	M00039|V$CREB_01, M00040\|V$CREBP1_01, M00041|V$CREBP1CJUN_01, M00113\|V$CREB_02, M00177|V$CREB_Q2, M00178\|V$CREB_Q4, M00916|V$CREB_Q2_01, M00917\|V$CREB_Q4_01, MA0018	(Herdegen and Leah 1998; Walton and Dragunow 2000)
	CREB,ATF	M00801|V$CREB_Q3, M00981\|V$CREBATF_Q6	(Walton and Dragunow 2000)
	E2F	M00803|V$E2F_Q2	(Dabrowski et al. 2006)
	Egr2	M00246|V$EGR2_01	(O'Donovan et al. 1999)
	Myb	MA0100	(Shin et al. 2001)
	Nfil3	M00045|V$E4BP4_01	(Junghans et al. 2004)
	NRSF	M00256|V$NRSF_01	(Schoenherr and Anderson 1995)
	Rfx5	M00626|V$EFC_Q6	(Blackshear et al. 2003; Durand et al. 2000)
	SMAD	M00974|V$SMAD_Q6_01	(Nakashima et al. 1999; Rodriguez et al. 2001)
	unknown	DME21|DGGVDRGAGSWR	
Heart	AP4	M00175|V$AP4_Q5	(Apone and Hauschka 1995; Flink et al. 1992; Smith and Lomax 1993)
	MEF2	M00232|V$MEF2_03	(Smith et al. 2005; Smith et al. 2007; Wasserman and Fickett 1998)
	Muscle TBP motif	M00320|V$MTATA_B	(Diagana et al. 1997)
	RORA	M00156|V$RORA1_01, MA0071	(Megy et al. 2002)
	Sf1	M00727|V$SF1_Q6	
	SRF	M00215|V$SRF_C	(Smith et al. 2005; Smith et al. 2007; Wasserman and Fickett 1998)
	unknown	DME10|SAGRRBAKRGRM, DME8|MVRGGRCAGR	
Kidney	HNF1	M00132|V$HNF1_01, M00790\|V$HNF1_Q6, M01011|V$HNF1_Q6_01, MA0046	(Senkel et al. 2005)
	Pax2	M00098|V$PAX2_01	(Schedl and Hastie 2000)
	unknown	DME11|SAKSKCTGKS	
Liver	Cutl1	M00104|V$CDPCR1_01	
	HNF4	MA0114	(Smith et al. 2005; Smith et al. 2007)
	PPAR, HNF-4, COUP, RAR	M00762|V$DR1_Q3	(Smith et al. 2005; Smith et al. 2007)
	unknown	DME27|WSDGARABSYWG	
mES c1	unknown	DME6|WABYCCWGMA	
mES c2	E2F1	M00940|V$E2F1_Q6_01	(Stead et al. 2002)
	Myc-Max	M00118|V$MYCMAX_01	(Takahashi and Yamanaka 2006)

Supplementary Data Tables

The following tables are available for download from:

http://bioinformatics-renlab.ucsd.edu/rentrac/wiki/MousePromoter

The site hosts information for the parallel publication submitted during the preparation of the dissertation. (Username: mouse, Password: 6tissues)

The Supplementary Data Tables are listed as "Data Tables" under the Supplementary Information section.

Data Table 4-1. 24,363 sites of Pol II binding annotated with known transcripts, CAGE, and Entrez Gene annotation as well as measures of tissue-specific Pol II binding using entropy (H) and categorical tissue-specificity (Q).

Data Table 4-2. Large regions of Pol II binding annotated with coordinates, tissue-enrichment, and matching Entrez Gene.

Data Table 4-3. TSS and CAGE-unmatched Pol II binding near Oct4 and Nanog binding sites.

Data Table 4:4-9: Tissue-specific gene promoter tables are annotated with coordinates, Entrez gene annotation, matching GO biological process and entropy measures based on Pol II binding and expression in various tissues:

Data Table 4-4. Brain

Data Table 4-5. Heart

Data Table 4-6. Kidney

Data Table 4-7. Liver

Data Table 4-8. mES c1

Data Table 4-9. mES c2

Data Table 4-10. 27 Refseq promoters tested by Pol II ChIP-qPCR in mES.

Data Table 4-11. Coordinates of 29 sites tested by Pol II ChIP-qPCR in liver.

Data Table 4-12. Coordinates of 5 randomly selected promoters with variable Pol II binding tested for Pol II binding and H3K4me3 by ChIP-qPCR in brain, heart, kidney, liver, and mES.

Chapter 5

Conclusions

In this dissertation, I have discussed the genome-wide mapping and analysis of active promoters, in human fibroblast cells and across a panel of adult mouse organs and embryonic stem cells. The scale and novelty of these studies required key collaborations with experimental scientists who performed the ChIP-chip experiments and biological validation, technology developers who provided the tiling array platforms, and statisticians with whom we collaborated to develop a model-based approach and efficient algorithm to identify binding sites from ChIP-chip data. As the bioinformatics researcher performing the bulk of the data management and analysis at the nexus of these collaborations, I have been pulled in the various directions of learning and performing ChIP-chip experiments, exploiting computational strategies to best tease out biological insights from our genome-wide promoter mapping data, and immersing myself in literature regarding transcription regulation and tissue-specific expression. It has been a challenge and a privilege. In this final chapter, I briefly review persisting ChIP-chip analysis issues beyond the scope of this dissertation and broadly outline some biological questions brought to light by the work described.

5.1 ChIP-chip Analysis Issues

"Much of what we present in this chapter could be described as first pass attempts to deal with the deluge of data arriving at our doors. Questions come in a volume and

pace that demand answers; we do not have the luxury of waiting until we have final solutions to problems…"[196]

- Yee Hwa Yang and Terry Speed, *Statistical Analysis of Gene Expression Data*

The quote above can be applied to this dissertation and ChIP-chip tiling array analysis to date. Although we have described advances in developing model-based approaches for analyzing ChIP-chip data and adapting microarray pre-processing strategies, improvements in analysis remain to be achieved. Many key issues with ChIP-chip data analysis are reminiscent of fundamental issues that have plagued microarray expression profiling. More than a decade since the first publications showing the utility of microarray expression profiling, experimental and analysis issues are still being reviewed and debated [197-199]. The following list highlights some of these microarray issues as adapted to ChIP-chip:

(1) Experimental design

(2) Quality control of ChIP-chip experiments

    a. Microarray quality

    b. Sample quality

        i. Antibody

        ii. Chromatin

    c. Hybridization

(3) Data pre-processing

    a. Image acquisition

    b. Normalization

(4) Semi-quantitative ChIP-chip enrichment values

a. Signal-to-noise ratio

b. Small dynamic range (ChIP-chip $\log_2$ratios compared to ChIP with quantitative PCR fold differences)

c. Accuracy

(5) Insufficient models for sources of variation

a. Probe-specific effects

(6) Experimental assessment of sensitivity and specificity

a. Comparison against ChIP with quantitative PCR for random and selected sites.

For instance, the question of accuracy (4c) or "conformity of a measured quantity to its actual value" has rarely been addressed in ChIP-chip experiments. Microarray expression signals attempt to model the number of transcripts or relative transcript levels in a cell [196,197]. ChIP-chip enrichment values attempt to represent protein-DNA interactions. Are we measuring protein-DNA interactions comparable to the traditional measures of binding affinity such as dissociation constants? If not, then what are we measuring? How does the duration of formaldehyde cross-linking affect what we measure? Even before the advent of ChIP-chip, cross-linking time has been judged to be a critical parameter in chromatin immunoprecipitation experiments that can affect antigen availability in chromatin and amount of total material. For instance, for proteins other than histones, longer cross-linking times are generally recommended [39].

Researchers analyzing ChIP-chip data need a fundamental understanding of the experimental procedure to derive realistic models of the data and the biological variation being measured. Aside from cross-linking time, there are still steps in the ChIP-chip

procedure for which variable methods are employed. It is still not clear how these different methods result in different ChIP-chip results. For instance, when mapping histone modifications in a genome, the use of sonication versus micrococcal nuclease digestion during the fragmentation step changes the profile of ChIP enrichment. In addition, there are at least three methods for amplifying IP-enriched DNA to generate enough sample for array hybridization -- linear amplification, whole genome amplification (WGA), and ligation-mediated PCR (LM-PCR) [200-202]. A systematic comparison of these amplification methods across various types of ChIP-chip tiling experiments – histone modifications, sequence-specific binding factors, and general transcriptional machinery – might reveal biases of the different approaches.

Relative to gene-centric microarray expression profiling data, ChIP-chip tiling array data differs in that there is no clear fundamental unit of probe or probe sets which correspond to a genomic site of binding or association for a particular factor. Typically the range of binding is inferred from the range of enriched signal above a selected background or following a model of enrichment. This difference makes ChIP-chip data simultaneously more challenging and more informative. It is more challenging because there is no clear fundamental unit for comparison across experiments or for summarizing binding sites. Binding sites can be provided as probe-based peaks or as variable genomic ranges. On the other hand it can be more informative because aside from large-scale discovery of binding sites, ChIP-chip can reveal patterns and profiles of binding revealing domains of gene regulation [139,190]. For instance, large domains of H3K27 methylation were observed at highly conserved genomic regions encompassing genes for developmental regulators in mouse embryonic stem cells. Within large domains of this

histone modification linked to repression, punctate sites of H3K4me3 were found at promoters and predicted to poise developmental regulators for transcription [139,190].

Clearly, interesting discoveries are being made using ChIP-chip as revealed by this dissertation and the onslaught of publications in the field. However, without a doubt, more work remains to be done to improve how we model and analyze ChIP-chip data. Furthermore, despite the current emphasis on its use for genomic annotation of binding sites, ChIP-chip is increasingly used to map a factor across cell types and conditions to examine changes in occupancy. Although we have taken a first step by adapting information theoretic measures such as entropy, strategies akin to differential gene expression analysis or gene-set expression analysis (GSEA) need to be developed for gleaning biological insight from changes in factor binding at genomic sites across conditions [203]. Finally, a broader question is how the emergence of cost-effective sequencing technologies will supplant the use of microarrays for identifying genomic sites of ChIP-enrichment [204]. Although ChIP-sequencing might render some array-based issues of probe design, image extraction, and normalization obsolete, some of the issues outlined above will remain pertinent.

5.2 Future Work

Although there are many conceivable questions emerging from our genome-wide mapping of promoters, I would like to highlight two areas for further investigation:

(1) Transcription elongation

(2) Active histone modifications at promoters

5.2.1 Transcription Elongation

Despite our emphasis on the importance of transcription initiation as a rate-limiting step in gene expression, our genome-wide mapping of active promoters by PIC binding in human fibroblast cells compared with microarray gene expression profiling data revealed a class of genes (15% of genes examined) with promoters bound by the PIC and no detectable transcripts (PIC Class II). This discordance can be trivially attributed to platform sensitivity differences, but more interestingly we have hypothesized that these genes might be regulated past the stage of transcription initiation – at the level of promoter clearance, elongation, or by post-transcriptional regulatory mechanisms such as mRNA degradation mediated by miRNAs (Chapter 3).

Increasingly, other groups have begun to emphasize barriers to transcription elongation at the various steps of promoter clearance, promoter-proximal pausing, and productive elongation [149]. Although Pol II at the stage of transcription initiation has been associated with Serine 5 phosphorylation at the CTD, the extent of this phosphorylation might not be sufficient to distinguish it from the hypo-phosphorylated form of Pol II we have been mapping as part of the PIC. Given that pausing has been shown to occur at sites +20 to +40 from the TSS, our ChIP-chip enrichment profiles do not give enough resolution to distinguish paused Pol II from Pol II at the initiation site. One group in particular has begun to examine some of the genes we have classified as Class II from our genome-wide mapping of active promoters. Using a nuclear run-on assay (NRO) to measure the density of transcriptionally engaged Pol II across a gene, they validated promoter proximal pausing at the 15 of 21 Class II genes they tested [205]. Interestingly, they also found evidence of promoter proximal pausing at half of the genes they tested

which we found to have PIC binding and detectable transcript (Class I) or no PIC binding and detectable transcript (Class II). To the best of their knowledge, only 7 human genes have been shown to have paused polymerase by NRO analysis until this recent work [205].

Clearly, the extent of regulation at the level of transcription elongation remains unclear, but our work has supported the possibility that this is greater than previously thought. The development of nuclear run-on analysis at a global scale might reveal the degree to which promoter-proximal pausing occurs. The extent to which this prevents productive elongation might also be investigated by mapping components of the known candidates for pausing control. These include DRB sensitivity-inducing factor (DSIF), negative elongation factor (NELF), as well as the positive transcription-elongation factor-b (P-TEFb) [149]. Given the current footprint of ChIP-chip enrichment, resolving the presence of pausing control factors, especially at short genes, might not be possible. However, preliminary tests of the binding resolution of components of the pausing control factors might be productive until global nuclear run-on assays are implemented.

5.2.3 Active Histone Modifications at Promoters

The presence of histone modifications associated with transcriptional activity, H3K4me3 and H3ac, at promoters with weak to undetectable Pol II binding and expressed transcript was a striking observation in our genome-wide promoter mapping across adult mouse organs and embryonic stem cells. Although we singled out a notable class of gene promoters which have enriched Pol II binding and chromatin modifications in embryonic stem cells and maintain the active chromatin marks without the Pol II binding in the differentiated tissues, our tests at 5 random promoters revealed the general

persistence of active chromatin marks, in particular H3K4me3, at promoters in tissues where the Pol II binding is undetectable or below the threshold determined by ChIP-qPCR. Our work was initially focused on the identification and characterization of "tissue-specific promoters" across mouse brain, heart, kidney, liver, and mES cells by Pol II binding. However, our comparison of the chromatin modifications at the Pol II sites across the tissues revealed our limited understanding for the establishment and the persistence of these chromatin marks, especially H3K4me3, which by consensus, have been linked to transcriptional activity [189].

A limited study at a handful of genes in human hepatic cell lines revealed the persistence of histone modifications – H3K4me2, H3K4me3, H3K79me2, H3ac, and H4ac – for an extended period of time after alpha-amanitin induced transcriptional block and through mitotic cell division [188]. The theory of histone modifications as "short-term memory of recent transcription" was first put forth based on observations of H3K4me3 enrichment at 5' ends of yeast genes that have been transcriptionally inactivated [206]. Recently, a study of DNA methylation, Pol II occupancy, and H3K4me2 at 16,000 human promoters revealed the enrichment of H3K4me2 at transcriptionally inactive CpG island promoters without DNA methylation. They suggest that the association of this active chromatin modification, H3K4me2, at transcriptionally inactive CpG islands might function in protecting the associated promoters from DNA methylation and silencing [192].

A recent review of H3K4me3 highlighted the potential complexity of this individual chromatin modification in humans by listing an extended family of histone methyltransferases which can establish this one mark in humans compared to a single enzyme in yeast. Similarly, they presented two superfamilies of candidate "effectors" or

factors with characteristic folds which have been shown to recognize the H3K4 methyl-lysines [189]. A better understanding of the specialized functions of the diverse writers and readers of this particular mark which has been shown to be, on its own, predictive of promoter 5' ends, might also be revealing of different modes of transcriptional regulation that regulate the expression of different classes of genes.

To the extent that we can continue to take advantage of the power of ChIP-chip, several future experiments might clarify some of our observations. For instance, how does the genomic localization and enrichment of H3K4me3 change across a time-course of embryonic stem-cell differentiation? If we complement these H3K4me3 marks with maps of candidate writers and readers of this mark, we can further ask the question of which writers and readers are gained and/or lost at genes depending on whether they have been transcriptionally inactivated or transcriptionally induced in the time-course. We can examine any biases of ChIP-chip enrichment of these various factors at CpG Island versus non-Cpg Island promoters. Without a doubt, elucidating the "intricacy" of a single mark, H3K4me3, at a specific class of regulatory elements, promoters, presents a formidable challenge on its own [189]. We hope that clever use of insights from our studies, as well as the work of many others, will pave the way to unraveling its complexity.

Bibliography

1. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;420:520-62.

2. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431:931-45.

3. Felsenfeld G, Groudine M. Controlling the double helix. *Nature* 2003;421:448-53.

4. Maniatis T, Goodbourn S, Fischer JA. Regulation of inducible and tissue-specific gene expression. *Science* 1987;236:1237-45.

5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.

6. 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.

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

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

9. Maire P, Wuarin J, Schibler U. The role of cis-acting promoter elements in tissue-specific albumin gene expression. *Science* 1989;244:343-6.

10. Li Q, Harju S, Peterson KR. Locus control regions: coming of age at a decade plus. *Trends Genet* 1999;15:403-8.

11. Carey MMF. Transcriptional regulation in eukaryotes: concepts, strategies, and techniques. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, c2000.

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

13. Malik S, Roeder RG. Dynamic regulation of pol II transcription by the mammalian Mediator complex. *Trends Biochem Sci* 2005;30:256-63.

14. Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway JW. The mammalian Mediator complex and its role in transcriptional regulation. *Trends Biochem Sci* 2005;30:250-5.

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

16. 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.

17. Reinberg D, Orphanides G, Ebright R, Akoulitchev S, Carcamo J, Cho H, Cortes P, Drapkin R, Flores O, Ha I, Inostroza JA, Kim S, Kim TK, Kumar P, Lagrange T, LeRoy G, Lu H, Ma DM, Maldonado E, Merino A, Mermelstein F, Olave I, Sheldon M, Shiekhattar R, Zawel L, et al. The RNA polymerase II general transcription factors: past, present, and future. *Cold Spring Harb Symp Quant Biol* 1998;63:83-103.

18. 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.

19. 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.

20. Levine M, Rubin GM, Tjian R. Human DNA sequences homologous to a protein coding region conserved between homeotic genes of Drosophila. *Cell* 1984;38:667-73.

21. 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.

22. Landry JR, Mager DL, Wilhelm BT. Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet* 2003;19:640-8.

23. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. *Genome Res* 2004;14:1711-8.

24. 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.

25. Smith AD, Sumazin P, Zhang MQ. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A* 2005;102:1560-5.

26. FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C. Clustering of DNA sequences in human promoters. *Genome Res* 2004;14:1562-74.

27. 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.

28. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the

mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 2004;101:6062-7.

29. 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.

30. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res* 2003;13:1290-300.

31. Timmusk T, Palm K, Metsis M, Reintam T, Paalme V, Saarma M, Persson H. Multiple promoters direct tissue-specific expression of the rat BDNF gene. *Neuron* 1993;10:475-89.

32. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell. New York: Garland Publishing, 2002.

33. Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T. Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat Cell Biol* 2004;6:73-7.

34. Henikoff S, Ahmad K. Assembly Of Variant Histones Into Chromatin. *Annu Rev Cell Dev Biol* 2005;21:133-153.

35. Wolffe AP. Inheritance of chromatin states. *Dev Genet* 1994;15:463-70.

36. Jenuwein T, Allis CD. Translating the histone code. *Science* 2001;293:1074-80.

37. van Steensel B, Henikoff S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* 2000;18:424-8.

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

39. Orlando V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* 2000;25:99-104.

40. 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.

41. 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.

42. 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.

43. 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.

44. Kirmizis A, Bartley SM, Kuzmichev A, Margueron R, Reinberg D, Green R, Farnham PJ. Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev* 2004;18:1592-605.

45. 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.

46. 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.

47. 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.

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

49. 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.

50. 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.

51. Kurdistani SK, Tavazoie S, Grunstein M. Mapping global histone acetylation patterns to gene expression. *Cell* 2004;117:721-33.

52. 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.

53. Roh TY, Ngau WC, Cui K, Landsman D, Zhao K. High-resolution genome-wide mapping of histone modifications. *Nat Biotechnol* 2004;22:1013-6.

54. Liang G, Lin JC, Wei V, Yoo C, Cheng JC, Nguyen CT, Weisenberger DJ, Egger G, Takai D, Gonzales FA, Jones PA. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc Natl Acad Sci U S A* 2004;101:7357-62.

55. 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.

56. 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.

57. Costa RH, Kalinichenko VV, Holterman AX, Wang X. Transcription factors in liver development, differentiation, and regeneration. *Hepatology* 2003;38:1331-47.

58. 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.

59. 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.

60. 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.

61. Lieb JD, Clarke ND. Control of Transcription through Intragenic Patterns of Nucleosome Composition. *Cell* 2005;123:1187-90.

62. 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.

63. 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.

64. 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.

65. 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.

66. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* 2001;106:697-708.

67. Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* 2004;83:349-60.

68. Buck MJ, Nobel AB, Lieb JD. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol* 2005;6:R97.

69. Isogai Y, Takada S, Tjian R, Keles S. Novel TRF1/BRF target genes revealed by genome-wide analysis of Drosophila Pol III transcription. *Embo J* 2007;26:79-89.

70. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A* 2006;103:12457-62.

71. Qi Y, Rolfe A, MacIsaac KD, Gerber GK, Pokholok D, Zeitlinger J, Danford T, Dowell RD, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. High-resolution computational models of genome binding events. *Nat Biotechnol* 2006;24:963-70.

72. Zheng M, Barrera LO, Ren B, Wu Y. ChIP-chip: Data, Model, and Analysis Proceedings of the American Statistical Association, Statistical Computing Section. Alexandria, VA: American Statistical Association., 2005.

73. Grant G, Manduchi,E, Stoeckert Jr,CJ. Analysis and Management of Microarray Gene Expression Data Current Protocols in Molecular Biology: John Wiley & Sons, 2006.

74. 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.

75. 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.

76. Kim TH, Ren B. Genome-wide analysis of Protein-DNA interactions. *Annual Review of Genomics and Human Genetics* 2006;7.

77. 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.

78. Kuzmichev A, Margueron R, Vaquero A, Preissner TS, Scher M, Kirmizis A, Ouyang X, Brockdorff N, Abate-Shen C, Farnham P, Reinberg D. Composition and histone substrates of polycomb repressive group complexes change during cellular differentiation. *Proc Natl Acad Sci U S A* 2005;102:1859-64.

79. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* 2002;16:245-56.

80. Weinmann AS, Farnham PJ. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* 2002;26:37-47.

81. 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:in press.

82. Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol* 1999;17:974-8.

83. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-73.

84. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002;30:e15.

85. Smyth GK. Limma: linear models for microarray data. In: R. Gentleman VC, S. Dudoit, R. Irizarry, W. Huber, ed. Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer, 2005:397-420.

86. Cleveland WS. Lowess: Robust Locally Weighted Regression for Smoothing and Graphing Data in Two or More Dimensions Association for Computing MachineryL:SIGGRAPH, 1983.

87. Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 2002;32 Suppl:490-5.

88. Wu Z, Irizarry RA. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol* 2005;12:882-93.

89. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;19:185-93.

90. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. *Cell* 2000;102:109-26.

91. Efron B. Local False Discovery Rates. Stanford: Stanford University, 2005:30.

92. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. Global identification of human transcribed sequences with genome tiling arrays. *Science* 2004;306:2242-6.

93. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006;34:D590-8.

94. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005;33:D54-8.

95. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003;4:P3.

96. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 2005;23:137-44.

97. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25-9.

98. Ji X, Li W, Song J, Wei L, Liu XS. CEAS: cis-regulatory element annotation system. *Nucleic Acids Res* 2006;34:W551-4.

99. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. *Bioinformatics* 2004;20:1453-4.

100. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998;95:14863-8.

101. Saldanha AJ. Java Treeview--extensible visualization of microarray data. *Bioinformatics* 2004;20:3246-8.

102. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;29:365-71.

103. Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ, Jr., Brazma A. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* 2002;3:RESEARCH0046.

104. Tjian R, Maniatis T. Transcriptional activation: a complex puzzle with few easy pieces. *Cell* 1994;77:5-8.

105. Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM. Identification and functional analysis of human transcriptional promoters. *Genome Res* 2003;13:308-12.

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

107. Ruppert S, Wang EH, Tjian R. Cloning and expression of human TAFII250: a TBP-associated factor implicated in cell-cycle regulation. *Nature* 1993;362:175-9.

108. Suzuki Y, Yamashita R, Sugano S, Nakai K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res* 2004;32 Database issue:D78-81.

109. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* 2003;31:34-7.

110. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. *Nucleic Acids Res* 2004;32 Database issue:D23-6.

111. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T. Ensembl 2004. *Nucleic Acids Res* 2004;32 Database issue:D468-70.

112. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* 1993;90:11995-9.

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

114. Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res* 2004;32 Database issue:D109-11.

115. Consortium IHGS. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431:931-45.

116. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 2004;14:331-42.

117. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. Using the transcriptome to annotate the genome. *Nat Biotechnol* 2002;20:508-12.

118. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. The transcriptional activity of human Chromosome 22. *Genes Dev* 2003;17:529-40.

119. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch

JB. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* 2002;99:4465-70.

120. Spellman PT, Rubin GM. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol* 2002;1:5.

121. Roy PJ, Stuart JM, Lund J, Kim SK. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. *Nature* 2002;418:975-9.

122. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 2001;291:1289-92.

123. Maniatis T, Reed R. An extensive network of coupling among gene expression machines. *Nature* 2002;416:499-506.

124. Krumm A, Hickey LB, Groudine M. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev* 1995;9:559-72.

125. Ambros V. The functions of animal microRNAs. *Nature* 2004;431:350-5.

126. Yuh CH, Bolouri H, Davidson EH. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 1998;279:1896-902.

127. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistics, UC Berkeley, Tech Report*;578.

128. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000;7:203-14.

129. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. *Science* 2004:1103388.

130. 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.

131. 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.

132. 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.

133. Sharov AA, Piao Y, Matoba R, Dudekula DB, Qian Y, VanBuren V, Falco G, Martin PR, Stagg CA, Bassey UC, Wang Y, Carter MG, Hamatani T, Aiba K, Akutsu H, Sharova L, Tanaka TS, Kimber WL, Yoshikawa T, Jaradat SA, Pantano S, Nagaraja R, Boheler KR, Taub D, Hodes RJ, Longo DL, Schlessinger D, Keller J, Klotz E, Kelsoe G, Umezawa A, Vescovi AL, Rossant J, Kunath T, Hogan BL, Curci A, D'Urso M, Kelso J, Hide W, Ko MS. Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biol* 2003;1:E74.

134. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, al e. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 2002;420:563-73.

135. Levine M, Tjian R. Transcription regulation and animal diversity. *Nature* 2003;424:147-51.

136. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng WT, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR. The functional landscape of mouse gene expression. *J Biol* 2004;3:21.

137. 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;38:626-35.

138. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, al e. The transcriptional landscape of the mammalian genome. *Science* 2005;309:1559-63.

139. 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;39:311-318.

140. 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.

141. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I. Predicting tissue-specific enhancers in the human genome. *Genome Res* 2007;17:201-11.

142. 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.

143. Smith AD, Sumazin P, Zhang MQ. Tissue-specific regulatory elements in mammalian promoters. *Mol Syst Biol* 2007;3:73.

144. Smith AD, Sumazin P, Xuan Z, Zhang MQ. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A* 2006;103:6275-80.

145. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* 2000;26:225-8.

146. Wasserman WW, Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 1998;278:167-81.

147. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2007;35:D5-12.

148. 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.

149. Saunders A, Core LJ, Lis JT. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol* 2006;7:557-67.

150. Cheng C, Sharp PA. RNA polymerase II accumulation in the promoter-proximal region of the dihydrofolate reductase and gamma-actin genes. *Mol Cell Biol* 2003;23:1961-7.

151. 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.

152. 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.

153. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, Decoste C, Schafer X, Lun Y, Lemischka IR. Dissecting self-renewal in stem cells with RNA interference. *Nature* 2006.

154. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CW, Zhao XD, Chiu KP, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* 2006;38:431-40.

155. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science* 2002;295:1306-11.

156. 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.

157. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 2005;433:769-73.

158. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;34:D140-4.

159. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. *J Mol Biol* 1987;196:261-82.

160. Cross SH, Bird AP. CpG islands and genes. *Curr Opin Genet Dev* 1995;5:309-14.

161. Episkopou V. SOX2 functions in adult neural stem cells. *Trends Neurosci* 2005;28:219-21.

162. GO Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 2006;34:D322-6.

163. Patterson LT, Potter SS. Hox genes and kidney patterning. *Curr Opin Nephrol Hypertens* 2003;12:19-23.

164. 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.

165. O'Donovan KJ, Tourtellotte WG, Millbrandt J, Baraban JM. The EGR family of transcription-regulatory factors: progress at the interface of molecular and systems neuroscience. *Trends Neurosci* 1999;22:167-73.

166. Harvey RP. NK-2 homeobox genes and heart development. *Dev Biol* 1996;178:203-16.

167. Plageman TF, Jr., Yutzey KE. T-box genes and heart development: putting the "T" in heart. *Dev Dyn* 2005;232:11-20.

168. Torban E, Goodyer P. What PAX genes do in the kidney. *Exp Nephrol* 1998;6:7-11.

169. Ruiz i Altaba A, Sanchez P, Dahmane N. Gli and hedgehog in cancer: tumours, embryos and stem cells. *Nat Rev Cancer* 2002;2:361-72.

170. Reya T, Clevers H. Wnt signalling in stem cells and cancer. *Nature* 2005;434:843-50.

171. Wei CL, Miura T, Robson P, Lim SK, Xu XQ, Lee MY, Gupta S, Stanton L, Luo Y, Schmitt J, Thies S, Wang W, Khrebtukova I, Zhou D, Liu ET, Ruan YJ, Rao M, Lim B. Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. *Stem Cells* 2005;23:166-85.

172. Niakan KK, Davis EC, Clipsham RC, Jiang M, Dehart DB, Sulik KK, McCabe ER. Novel role for the orphan nuclear receptor Dax1 in embryogenesis, different from steroidogenesis. *Mol Genet Metab* 2006;88:261-71.

173. Besser D. Expression of nodal, lefty-a, and lefty-B in undifferentiated human embryonic stem cells requires activation of Smad2/3. *J Biol Chem* 2004;279:45076-84.

174. Tanaka S, Kunath T, Hadjantonakis AK, Nagy A, Rossant J. Promotion of trophoblast stem cell proliferation by FGF4. *Science* 1998;282:2072-5.

175. Joyner AL, Liu A, Millet S. Otx2, Gbx2 and Fgf8 interact to position and maintain a mid-hindbrain organizer. *Curr Opin Cell Biol* 2000;12:736-41.

176. Kioussi C, Briata P, Baek SH, Wynshaw-Boris A, Rose DW, Rosenfeld MG. Pitx genes during cardiovascular development. *Cold Spring Harb Symp Quant Biol* 2002;67:81-7.

177. Alfano G, Vitiello C, Caccioppoli C, Caramico T, Carola A, Szego MJ, McInnes RR, Auricchio A, Banfi S. Natural antisense transcripts associated with genes involved in eye development. *Hum Mol Genet* 2005;14:913-23.

178. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006;34:D108-10.

179. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;32:D91-4.

180. Wang Z, Wang DZ, Pipes GC, Olson EN. Myocardin is a master regulator of smooth muscle gene expression. *Proc Natl Acad Sci U S A* 2003;100:7129-34.

181. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006;126:663-76.

182. Walton MR, Dragunow I. Is CREB a key to neuronal survival? *Trends Neurosci* 2000;23:48-53.

183. Durand B, Vandaele C, Spencer D, Pantalacci S, Couble P. Cloning and characterization of dRFX, the Drosophila member of the RFX family of transcription factors. *Gene* 2000;246:285-93.

184. Lau P, Nixon SJ, Parton RG, Muscat GE. RORalpha regulates the expression of genes involved in lipid homeostasis in skeletal muscle cells: caveolin-3 and CPT-1 are direct targets of ROR. *J Biol Chem* 2004;279:36828-40.

185. Megy K, Audic S, Claverie JM. Heart-specific genes revealed by expressed sequence tag (EST) sampling. *Genome Biol* 2002;3:RESEARCH0074.

186. Maity SN, Golumbek PT, Karsenty G, de Crombrugghe B. Selective activation of transcription by a novel CCAAT binding factor. *Science* 1988;241:582-5.

187. Vanden Heuvel GB, Brantley JG, Alcalay NI, Sharma M, Kemeny G, Warolin J, Ledford AW, Pinson DM. Hepatomegaly in transgenic mice expressing the homeobox gene Cux-1. *Mol Carcinog* 2005;43:18-30.

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

189. Ruthenburg AJ, Allis CD, Wysocka J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol Cell* 2007;25:15-30.

190. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 2006;125:315-26.

191. Roh TY, Cuddapah S, Cui K, Zhao K. The genomic landscape of histone modifications in human T cells. *Proc Natl Acad Sci U S A* 2006;103:15782-7.

192. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 2007;39:457-66.

193. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;5:R80.

194. Gershenzon NI, Ioshikhes IP. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* 2005;21:1295-300.

195. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999;22:281-5.

196. Chipman H, Dudoit, S., Fridlyand, J., Hastie, T.J., Li, C., Speed, T, Tibshirani, R., Tseng, G.C., Wong, W.H., Yang, Y.H. Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC, 2003:240.

197. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* 2006;22:101-9.

198. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W. Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2005;2:345-50.

199. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-80.

200. Liu CL, Schreiber SL, Bernstein BE. Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* 2003;4:19.

201. O'Geen H, Nicolet CM, Blahnik K, Green R, Farnham PJ. Comparison of sample preparation methods for ChIP-chip assays. *Biotechniques* 2006;41:577-80.

202. Ren B, Dynlacht BD. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol* 2004;376:304-15.

203. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-50.

204. Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. *Nat Rev Genet* 2006;7:632-44.

205. Core LJ, Lis, J. T. Analysis of promoter proximal pausing in human cell lines Systems Biology: Global Regulation of Gene Expression. Cold Spring Harbor, New York, 2007.

206. 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.